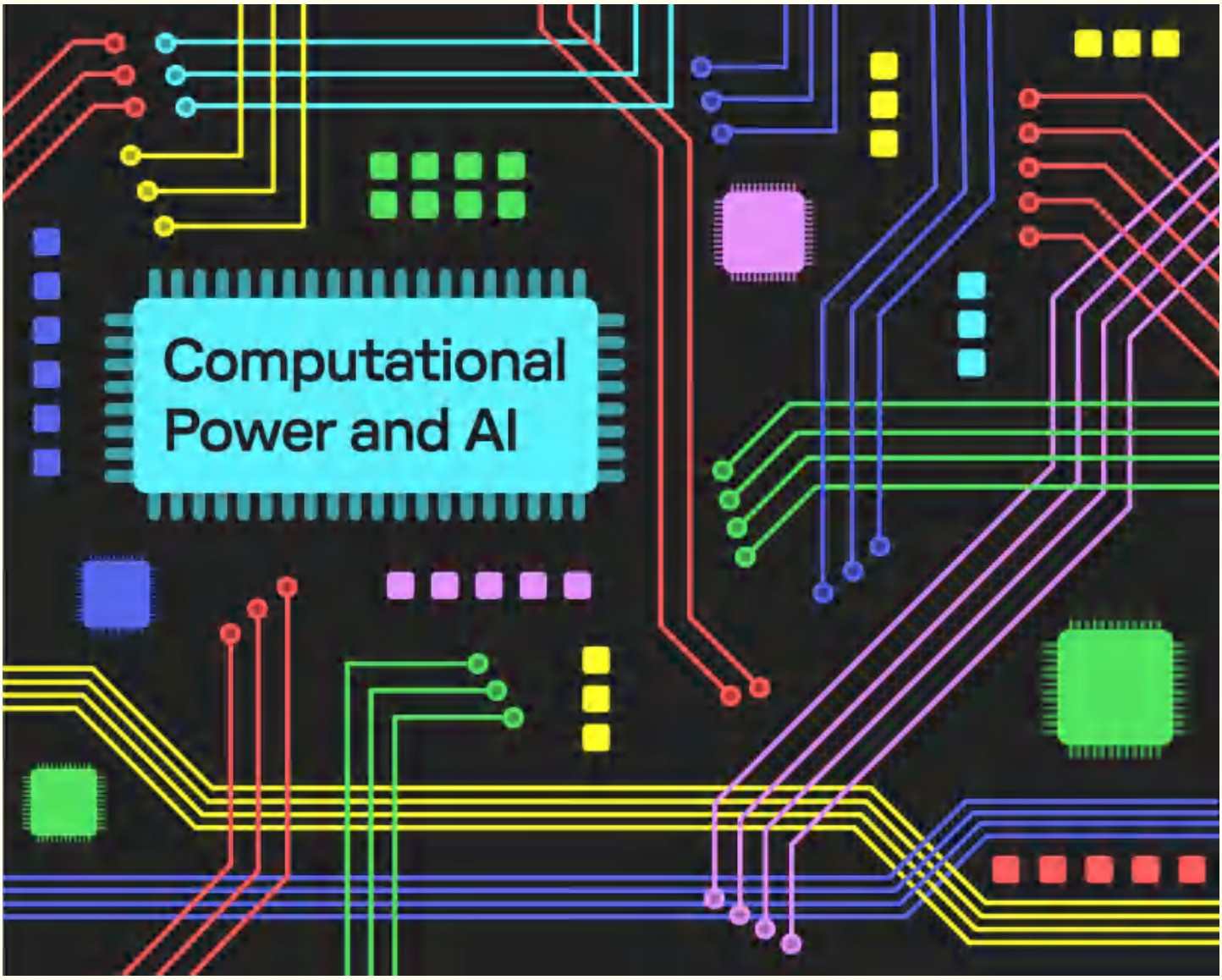
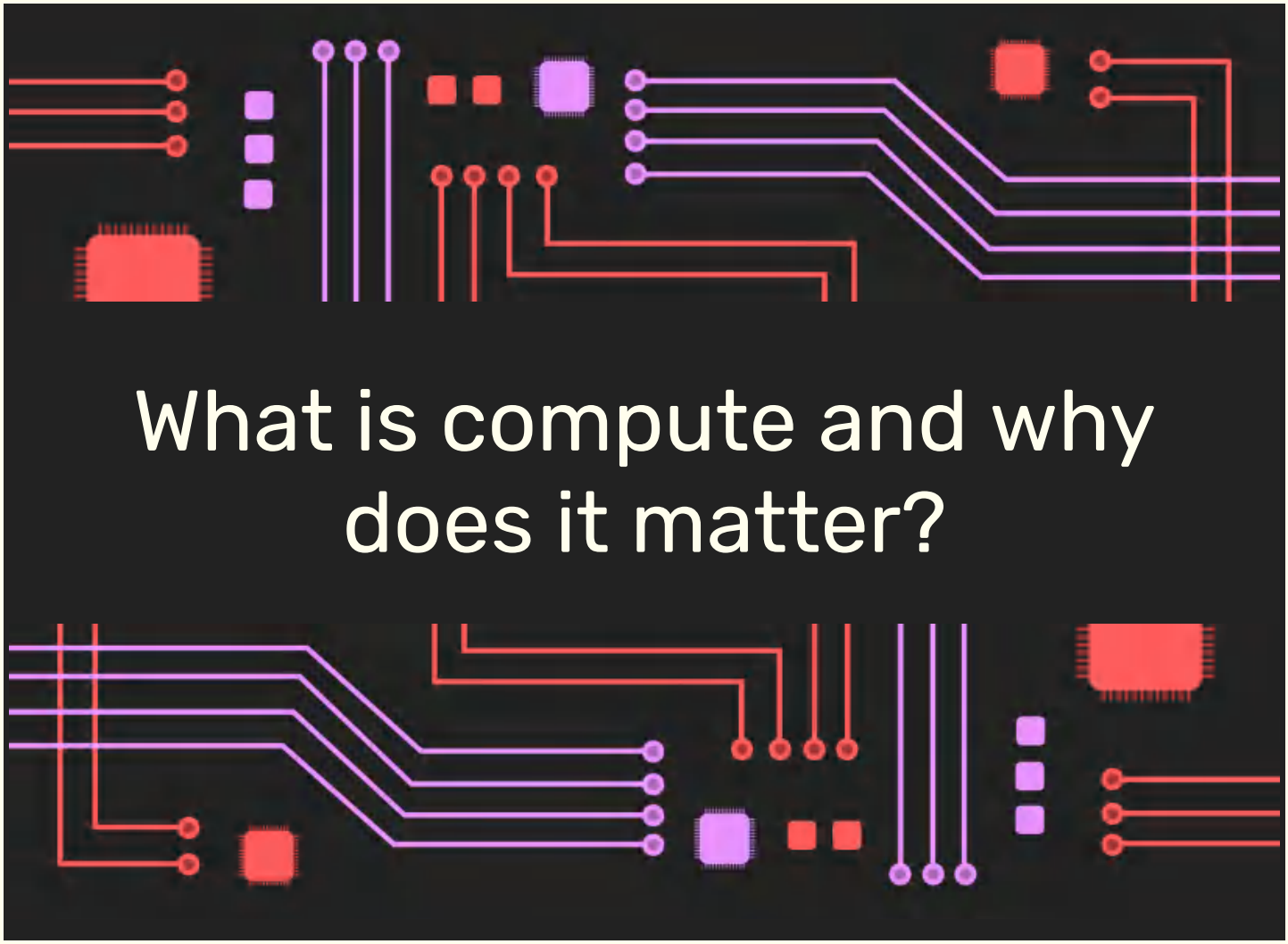
AINOW



By Jai Vipra + Sarah Myers West

AI Now Institute

Table of Contents



Computational power, or *compute*, is a core dependency in building large-scale AI.[1]

Amid a steadily growing push to build AI at larger and larger scale, access to compute—along with data and skilled labor—is a key component[2] in building artificial

[1] For those interested in a deeper dive, many other resources on compute power and AI provide a parallax view on these issues: see the researcher Mél Hogan's compilation of critical studies of the cloud; Seda Gürses's work on computational power and programmable infrastructures; Vili Lehdonvirta's work on cloud empires; and Nicole Starosielski's and Ingrid Burrington's material studies of networked infrastructures, among others.

[2] Amba Kak and Sarah Myers West, *AI* Now 2023 Landscape: *Executive Summary*, AI Now Institute, April 11, 2023, https://ainowinstitute.org/general/2023-landscape-executive-summary; Ben Buchanan, *The AI Triad and What It Means for National Security Strategy*, Center for Security and Emerging Technology (CSET), August 2020, https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy.

intelligence systems. It is profoundly monopolized at key points in the supply chain by one or a small handful of firms.[3]

Industry concentration acts as a shaping force in how computational power is manufactured and accessed by tech developers. As we will show, it influences the behavior of even the biggest AI firms as they encounter the effects of compute scarcity. A recent report from Andreessen Horowitz describes compute as "a predominant factor driving the industry today," noting that companies have spent "more than 80% of their total capital on compute resources."[4]

This concentration in compute also incentivizes cloud infrastructure providers to act in ways that will protect their dominant position in the market, racing to release products before they're ready for widespread use and behaving in ways that encourage lock-in into their cloud ecosystems.

Understanding the influence of computational infrastructure on the political economy of artificial intelligence is profoundly important: it affects who can build AI, what kind of AI gets built, and who profits along the way. It defines the contours of concentration in the tech industry, incentivizes toxic competition among AI firms,[5] and deeply impacts the environmental footprint of artificial intelligence.[6] It enables dominant firms to extract rents from consumers and small businesses dependent on their services, and creates systemic harms when systems fail or malfunction due to the creation of single points of failure. Most concerningly, it expands the economic and political power of the firms that have access to compute, cementing the control of firms that already dominate the tech industry.

Policy interventions—including industrial policy movements, export controls, and antitrust enforcement—likewise have a profound effect on who has access to compute, at what cost, and under what conditions. Thinking deliberately about policy mechanisms offers a path forward for mitigating the most harmful effects of AI. But many actors with divergent incentives are converging on compute as a leverage point for achieving their objectives: for example, cofounder of DeepMind Mustafa Suleyman recently called for sales of chips to be restricted to firms that can demonstrate compliance with safe and ethical uses of the technology,[7] while others are pointing to export controls to mitigate existential risk while

---

[3] Jai Vipra and Anton Korinek, "Market Concentration Implications of Foundation Models: The Invisible Hand of ChatGPT," *Brookings Institution, Sept. 7, 2023, https://www.brookings.edu/articles/market-concentration-implications-of-foundation-models-the-invisible-hand-of-chatgpt/*

[4] Guido Appenzeller, Matt Bornstein, and Martin Casado, *Navigating the High Cost of AI Compute*, Andreessen Horowitz, April 27, 2023, https://a16z.com/navigating-the-high-cost-of-ai-compute.

[5] AI Now Institute, *Toxic Competition: Regulating Big Tech's Data Advantage*, April 11, 2023, https://ainowinstitute.org/publication/toxic-competition; Sarah Myers West and Jai Vipra, *Computational Power and AI*, AI Now Institute, June 22, 2023, https://ainowinstitute.org/publication/policy/computational-power-and-ai.

[6] AI Now Institute, *The Climate Costs of Big Tech*, April 11, 2023, https://ainowinstitute.org/spotlight/climate.

[7] Richard Waters, "US Should Use Chip Leadership to Enforce AI Standards, Says Mustafa Suleyman," *Financial Times*, September 1, 2023, https://ft.com/content/f828fef3-862c-4022-99d0-41efbc73db80.

ignoring near-term harms.[8] Any such policy interventions will require careful calibration and thought, learning from the past decade of research and evidence on the implications of artificial intelligence, as well as looking to historical case studies in which measures such as nondiscrimination policy have been utilized to target firms' monopoly power over critical infrastructures.[9]

Understanding the material underpinnings of artificial intelligence is an important entry point for examining its effects on the broader public. This guide offers a primer for one key dimension: compute.

Defining "Compute"

When we use the word "compute," we sometimes mean the number of computations needed to perform a particular task, such as training an AI model. At other times, "compute" is used to refer solely to hardware, like chips. Often, though, we use "compute" to refer to a stack that includes both hardware and software. This stack can include

1. **chips**, such as Graphics Processing Units (GPUs), which we will examine in detail in a later section;
2. **software** to enable the use of specialized chips like GPUs;
3. **domain-specific languages** that can be optimized for machine learning;
4. **data management software**; and
5. **infrastructure** in data centers that allows the use of thousands of chips together, including cabling, servers, and cooling equipment.

The *amount* of compute used is measured in *floating point operations* (**FLOP**). In rough terms, a FLOP is a mathematical operation that enables the representation of extremely large numbers with greater precision. Compute *performance* is measured in *floating point operations per second* (**FLOP/s**), or how many computations a given resource can carry out in a second.

Recent progress in AI models has been made possible through **deep learning**, a machine learning technique that uses vast amounts of data to build layers of understanding. Deep learning has facilitated the development of models that have more generalized capabilities than we have seen before. It is enabled by the use of high-end computational resources that can perform many computations very quickly and in parallel.

---

[8] Gregory C Allen, Emily Benson and William Alan Reinsch, "Improved Export Controls Enforcement Technology Needed for National Security", *Center for Strategic & International Studies,* November 20, 2022, https://www.csis.org/analysis/improved-export-controls-enforcement-technology-needed-us-national-security.
[9] Barry C. Lynn, "The Big Tech Extortion Racket: How Google, Amazon, and Facebook Control Our Lives," *Harper's Magazine*, September 2020, https://harpers.org/archive/2020/09/the-big-tech-extortion-racket.

Deep learning is computationally expensive by design.[10] Researchers in AI have largely concluded that increasing scale is key to accuracy and performance in training deep learning models. This has driven an exponentially growing demand for computing power, leading to concerns that the current pace of growth is unsustainable.[11]

This trend has been borne out historically: before the deep learning era, the amount of compute used by AI models doubled in about 21.3 months; after deep learning as a paradigm took hold around 2010, the amount of compute used by models started doubling in only 5.7 months.[12] Since 2015, however, trends in compute growth have split in two: the amount of compute used in large-scale models has been doubling in roughly 9.9 months, while the amount of compute used in regular-scale models has been doubling in only about 5.7 months.[13]

The lack of sustainability cuts in two directions: first, there is a clear scarcity, particularly in the kinds of state-of-the-art (SOTA) chips needed for training large-scale AI models efficiently. Demand for these chips—currently Nvidia's H100 and A100—is extremely high. Supplies are limited, leading to unconventional arrangements such as the collateralization of GPUs to raise funds,[14] organizations set up to provide GPU rental services,[15] and purchases of GPUs by nation-states seeking a competitive advantage.[16] Demand for large amounts of compute power and the resulting scarcity are both products of public policy, and are a significant shaping force on the field's trajectory. This puts cloud infrastructure firms like Amazon Web Services (AWS), Google Cloud, and Microsoft Azure, as well as the chip design firm Nvidia and chip fabrication firm Taiwan Semiconductor Manufacturing Company (TSMC), in a dominant position.

Large-scale compute is also environmentally unsustainable: chips are highly toxic to produce[17] and require an enormous amount of energy to manufacture:[18] for example, TSMC

[10] Neil C. Thompson et al., "The Computational Limits of Deep Learning," arXiv, July 27, 2022, https://doi.org/10.48550/arXiv.2007.05558.

[11] Pengfei Li et al., "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models," arXiv, April 6, 2023, https://doi.org/10.48550/arXiv.2304.03271.

[12] Jaime Sevilla et al., "Compute Trends across Three Eras of Machine Learning," arXiv, March 9, 2022, https://doi.org/10.48550/arXiv.2202.05924.

[13] Ibid.

[14] Krystal Hu, "CoreWeave Raises $2.3 Billion in Debt Collateralized by Nvidia Chips," Reuters, August 3, 2023, https://reuters.com/technology/coreweave-raises-23-billion-debt-collateralized-by-nvidia-chips-2023-08-03.

[15] Erin Griffith, "The Desperate Hunt for the A.I. Boom's Most Indispensable Prize," *New York Times*, August 16, 2023, https://nytimes.com/2023/08/16/technology/ai-gpu-chips-shortage.html.

[16] James Titcomb, "Sunak to Spend £100m of Taxpayer Cash on AI Chips in Global Race for Computer Power," *Telegraph*, August 20, 2023, https://telegraph.co.uk/business/2023/08/20/sunak-spend-100m-taxpayer-cash-ai-chips-global-race.

[17] Pádraig Belton, "The Computer Chip Industry Has a Dirty Climate Secret," *Guardian*, September 18, 2021, https://theguardian.com/environment/2021/sep/18/semiconductor-silicon-chips-carbon-footprint-climate.

[18] "Can Computing Clean up Its Act?," *Economist*, August 16, 2023, https://economist.com/science-and-technology/2023/08/16/can-computing-clean-up-its-act.

on its own accounts for 4.8 percent of Taiwan's national energy consumption, more than the entire capital city of Taipei.[19] Running data centers is likewise environmentally very costly: estimates equate every prompt run on ChatGPT to the equivalent of pouring out an entire bottle of water.[20]

TSMC on its own accounts for 4.8 percent of Taiwan's national energy consumption, more than the entire capital city of Taipei.

Could future research directions lead to smaller models? To answer this question, it is helpful to look at why larger models took hold in the first place—and who benefits from perpetuating them. Sara Hooker's concept of the hardware lottery describes the phenomenon where a research idea wins because it is the most suited to the available hardware and software. In this sense, the hardware and software determine the research direction, not the other way around.[21] Deep neural networks at first represented an idea that was too ahead of its time in hardware terms, and was thus long ignored. It was only when the research on neural networks was combined with massive datasets scraped from the web, the computational resources accrued by Big Tech firms, and the incentive structures introduced by commercial surveillance that we saw the explosion of interest in building artificial intelligence systems.[22] Hooker predicts that due to increasing specialization in computing, the cost of straying from the mainstream, hardware-compatible set of ideas will only increase over time.

In other words, large models today are not only compatible with the hardware available today; they also provide returns to cloud infrastructure providers that have already made massive investments in their hardware. Given the high up-front costs of obtaining GPUs and networking, as well as of building the data center infrastructures needed to run compute at scale most efficiently, the players who own this infrastructure—hyperscalers like Google Cloud, Microsoft Azure, and Amazon Web Services—have strong incentives to

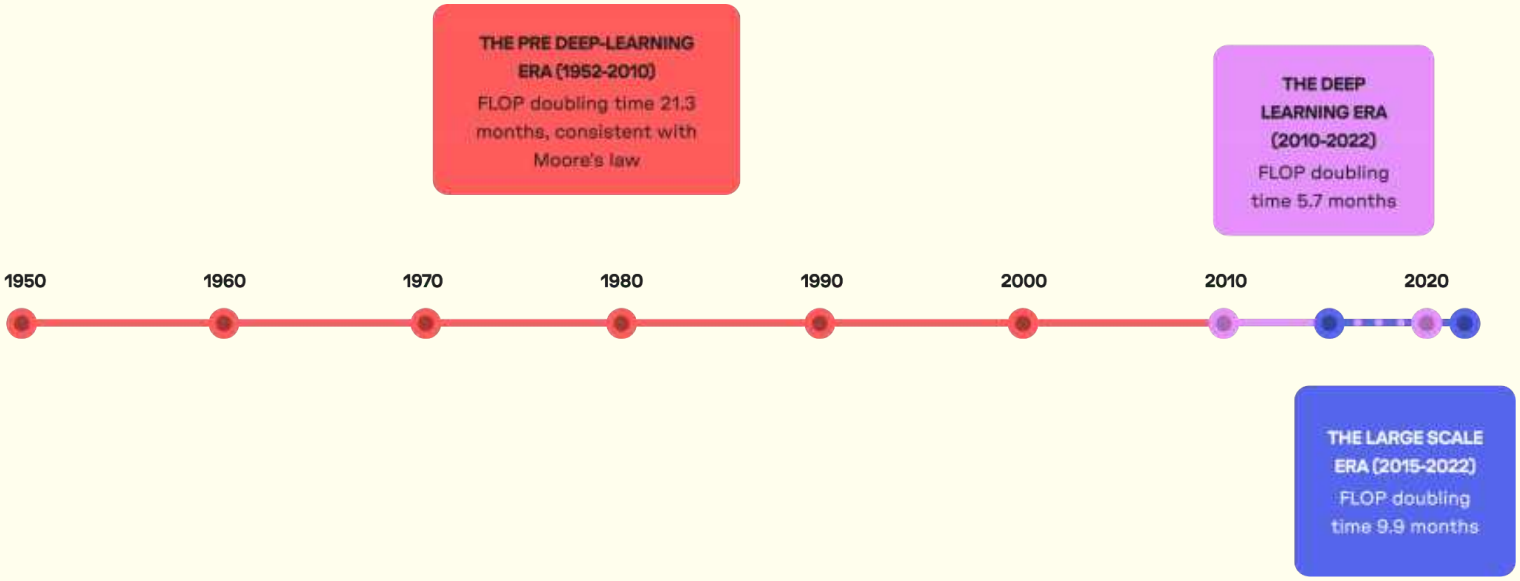
[19] Alynne Tsai, "TSMC's Push toward Green Energy," *Taipei Times,* July 17, 2020, https://taipeitimes.com/News/editorials/archives/2020/07/17/2003740051.

[20] Pengfei Li et al., "Making AI Less 'Thirsty.'"

[21] Sara Hooker, "The Hardware Lottery," *Communications of the ACM* 64, no. 12. (December 2021): 58–65, https://cacm.acm.org/magazines/2021/12/256929-the-hardware-lottery/fulltext.

[22] Emily Denton et al., "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet," *Big Data & Society* 8, no. 2 (July 1, 2021), https://doi.org/10.1177/20539517211035955; Meredith Whittaker, "The Steep Cost of Capture," *ACM Interactions,* accessed 9 September 2023, https://dl.acm.org/doi/10.1145/3488666.

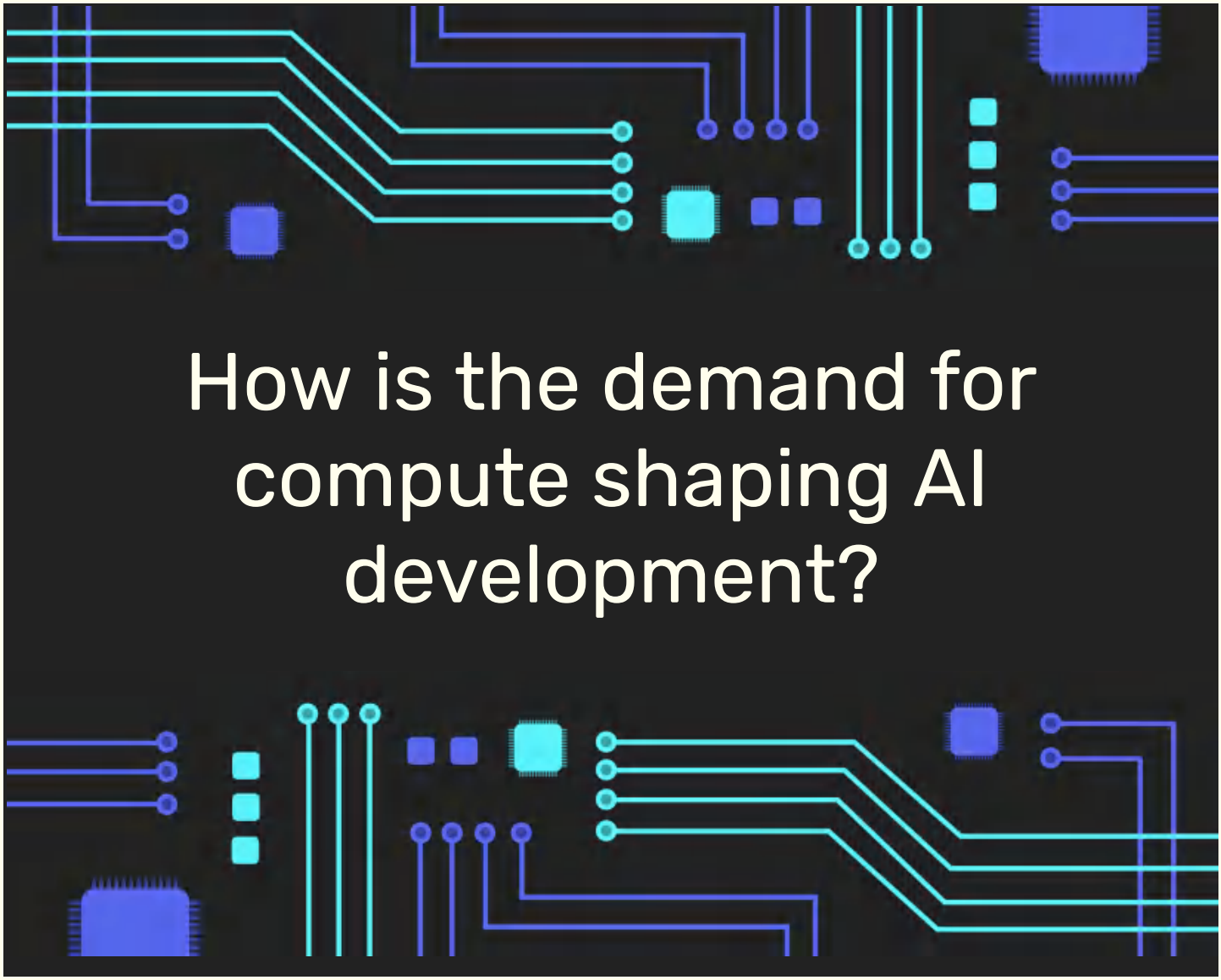
maximize these investments through behavior that seeks to perpetuate AI at scale, favors their ecosystem of corporate holdings, and that further locks in their dominance in cloud computing.[23] Hooker sees the most promise in interventions that target the software-hardware relationship,[24] while regulators around the globe are looking more deeply into concentration in the cloud ecosystem.[25]



[23] West and Vipra, "Computational Power and AI."

[24] Hooker, "The Hardware Lottery."

[25] FTC Office of Technology, "An Inquiry into Cloud Computing Business Practices: The Federal Trade Commission Is Seeking Public Comments," Federal Trade Commission, March 21, 2023, https://ftc.gov/policy/advocacy-research/tech-at-ftc/2023/03/inquiry-cloud-computing-business-practices-federal-trade-commission-seeking-public-comments; Ofcom, *Consultation: Cloud Services Market Study (Interim Report)*, July 7, 2023, https://ofcom.org.uk/consultations-and-statements/category-2/cloud-services-market-study; ACM, "ACM Launches Market Study into Cloud Services," May 18, 2021, https://acm.nl/en/publications/acm-launches-market-study-cloud-services; French Competition Authority, "The French Competition Authority Publishes Its Market Study on Competition in the Cloud Computing Sector," *E-Competitions Bulletin*, June 29, 2023, https://concurrences.com/en/bulletin/news-issues/june-2023/the-french-competition-authority-publishes-its-market-study-on-competition-in.



Compute is scarce, and has become a key bottleneck in both the training of large-scale AI models and the deployment of those models in AI products (often referred to in this literature as **"inference,"** or when the model is asked to generate a response). As Elon Musk told a room full of CEOs in May 2023, "GPUs are at this point considerably harder to get than drugs."[26]

This scarcity gives significant market power to the handful of companies that have amassed control over these resources—chip fabricators like the Taiwan Semiconductor Manufacturing Company; chip designers like Nvidia; and cloud infrastructure firms like Google, Microsoft, and Amazon—the largest of which also operate sprawling platform

[26] Deepa Seetharaman and Tom Dotan, "The AI Boom Runs on Chips, but It Can't Get Enough," *Wall Street Journal*, May 29, 2023, https://wsj.com/articles/the-ai-boom-runs-on-chips-but-it-cant-get-enough-9f76f554.

ecosystems that give them access to data and manifold ways to commercialize AI, and have first-mover advantage in large-scale artificial intelligence.[27]

Start-ups

The demand for compute holds firm even as new start-ups begin to build commercial AI products: to enter the field, small companies building models or making vertically integrated AI applications must secure compute credits or make other contractual arrangements with Big Tech firms. Outside of this, their options are to contract with providers of hosted model services like OpenAI and Hugging Face, which have partnerships with Microsoft and Amazon, respectively. Google has touted that 70 percent of generative AI startups use Google's cloud facilities, though those numbers should be taken with a grain of salt.[28] Building these resources from scratch is prohibitively expensive due to significant start-up costs, lack of interoperability at key points in the compute stack, and bottlenecks in the supply chain for key components of compute infrastructure. Talent requirements also grow as compute costs grow, because very specialized knowledge is needed to make the most of scarce hardware, and much of this knowledge is tacit.[29]

Google has touted that 70 percent of generative AI startups use Google's cloud facilities, though those numbers should be taken with a grain of salt.

Big Tech Firms

The effects of the compute bottleneck are evidenced in the behavior of even the largest AI firms: Microsoft, which runs its own Azure cloud infrastructure business, is now listing "availability of GPUs" as a risk factor in its annual reporting, noting that "our datacenters [sic] depend on the availability of permitted and buildable land, predictable energy,

[27] Richard Waters, "Adam Selipsky: There Will Not Be One Generative AI Model to Rule Them All," *Financial Times*, July 26, 2023, https://ft.com/content/5ffa06fa-28f8-47b2-b0c5-f46c1b14d5cd.

[28] Johan Moreno, "70% of Generative AI Startups Rely on Google Cloud, AI Capabilities," *Forbes*, July 25, 2023 https://forbes.com/sites/johanmoreno/2023/07/25/70-of-generative-ai-startups-rely-on-google-cloud-ai-capabilities-says-alphabet-ceo-sundar-pichai.

[29] Madhumita Murgia, "Big Tech Companies Use Cloud Computing Arms to Pursue Alliances with AI Groups," *Financial Times*, February 5, 2023, https://ft.com/content/5b17d011-8e0b-4ba1-bdca-4fbfdba10363.

networking supplies, and servers, including graphics processing units ('GPUs') and other components" and that the company "may have excessive outages, data losses, and disruptions of our online services if we fail to maintain an adequate operations infrastructure."[30] Microsoft is reportedly now rationing access to its hardware,[31] and considered a deal with Oracle to share AI servers to solve a GPU shortage.[32] Customers of the other major cloud infrastructure providers are likewise experiencing the strains of increased demand by encountering delays and scarcity of the most powerful GPUs,[33] further illustrating the detrimental effects of a highly concentrated industry.

Demand for compute and limited cash reserves following Elon Musk's departure from the company are what motivated OpenAI to convert from a nonprofit organization to a for-profit limited partnership. OpenAI CEO Sam Altman cites the need for large amounts of compute as a key shaping factor in many of the company's decisions, from system design to decisions about future products;[34] this drove Microsoft's $10 billion investment in OpenAI.[35] Altman told Congress: "We're so short on GPUs, the less people that use the tool, the better."[36] For its part, Google integrated its DeepMind and Google Brain teams under similar considerations,[37] and Nvidia's meteoric rise in market capitalization has also been driven by an understanding that the compute layer of AI systems is lucrative.[38] This point was driven home recently when CoreWeave raised $2.3 billion in debt from Magnetar and Blackstone by collateralizing its Nvidia H100 chips.[39]

Compute is scarce, expensive, and in demand even within the cloud infrastructure firms that have access to it. All of these factors incentivize the companies controlling key bottlenecks to deepen their dominant position over the market using whatever methods are at hand—including policy advocacy.

---

[30] United States Securities and Exchange Commission, Form 10-K for the Fiscal Year Ended June 2023, accessed September 9, 2023, https://sec.gov/Archives/edgar/data/789019/000095017023035122/msft-20230630.htm.

[31] Aaron Holmes and Kevin McLaughlin, "Microsoft Rations Access to AI Hardware for Internal Teams," *The Information*, March 15, 2023, https://theinformation.com/articles/microsoft-rations-access-to-ai-hardware-for-internal-teams.

[32] Aaron Holmes and Anissa Gardizy, "Microsoft and Oracle Discussed Sharing AI Servers to Solve Shortage," *The Information*, May 9, 2023, https://theinformation.com/articles/microsoft-and-oracle-discussed-sharing-ai-servers-to-solve-shortage.

[33] Appenzeller, Bornstein, and Casado, "Navigating the High Cost of AI Compute."

[34] Raza Habib, "OpenAI's Plans According to Sam Altman," Humanloop, May 28, 2023, accessed September 9, 2023, https://web.archive.org/web/20230601000258/https://website-nm4keew22-humanloopml.vercel.app/blog/openai-plans.

[35] Dina Bass, "Microsoft to Invest $10 Billion in ChatGPT Maker OpenAI," Bloomberg, January 23, 2023, https://bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai.

[36] Brian Fung, "The Big Bottleneck for AI: A Shortage of Powerful Chips," CNN Business, August 6, 2023, https://cnn.com/2023/08/06/tech/ai-chips-supply-chain/index.html.

[37] Google DeepMind, "Announcing Google DeepMind," April 20, 2023, https://deepmind.com/blog/announcing-google-deepmind; SA Transcripts, "Alphabet Inc. (GOOG) Q1 2023 Earnings Call Transcript," *Seeking Alpha* (blog), April 25, 2023, https://seekingalpha.com/article/4596558-alphabet-inc-goog-q1-2023-earnings-call-transcript.

[38] Gerrit De Vynck, "Why Nvidia Is Suddenly One of the Most Valuable Companies in the World," *Washington Post*, June 2, 2023, https://washingtonpost.com/technology/2023/05/25/nvidia-ai-stock-gpu-chatbots.

[39] Hu, "CoreWeave Raises $2.3 Billion in Debt Collateralized by Nvidia Chips."

Geopolitics

Compute scarcity is central to the emerging geopolitical landscape around AI. It is being used by countries as a core support in their industrial policy ambitions and as a retaliatory measure aimed at curbing the advancement of adversaries. Companies that operate key chokepoints like Nvidia, TSMC, and ASML are particularly exposed to the geopolitical dynamics of the so-called "AI Arms Race."

The semiconductor industry formed the heart of what became Silicon Valley. Forged in the late 1960s, the industry grew to become a core component of the US political economy,[40] referred to in a 2003 National Academies of Science Report as "the premier general-purpose technology of our post-industrial era."[41] US industrial policy has historically worked to support US dominance in semiconductor manufacturing through measures including diplomatic pressure, tolerance of industry bottlenecks, and state subsidies[42]—grounded in the perspective of industrial policy proponents like Robert Reich and Joseph Stiglitz, who believed the tech industry would inherently accrue "natural" monopolies.[43] But in the 1990s and 2000s, the US government reduced its level of investment as interventionist policy fell out of vogue. This set the stage for Intel's decline relative to firms like Taiwan Semiconductor Manufacturing Company, now the world's dominant chip fabricator; and ASML, a Dutch company that is the sole manufacturer of the equipment needed to build state-of-the-art chips.

This context forms the background of current geopolitical conflicts around compute power. The so-called "US-China AI Arms Race" places compute front and center: US export controls impose restrictions on private entities in the semiconductor sector in order to contain and degrade China's AI progress, restraining the country's access to state-of-the-art compute.[44] These export restrictions cover significant portions of the entire supply chain for chip manufacturing, not just the chips themselves.[45]

For example, chip design firm Nvidia is currently barred from selling its top-line A100 and H100 chips in China and can only offer downgraded A800 and H800 chips.[46] National

[40] Susannah Glickman, "Semi-Politics," *Phenomenal World* (blog), June 24, 2023, https://phenomenalworld.org/analysis/semi-politics.

[41] Ibid.

[42] Ibid.

[43] Ibid.

[44] Gregory C. Allen, "Choking Off China's Access to the Future of AI," Center for Strategic and International Studies (CSIS), October 11, 2022, https://csis.org/analysis/choking-chinas-access-future-ai.

[45] Thanks to Josh Lund for this point.

[46] Ana Swanson, "The Biden Administration Is Weighing Further Controls on Chinese Technology," *New York Times*, October 27, 2022, https://nytimes.com/2022/10/27/business/the-biden-administration-is-weighing-further-controls-on-chinese-technology.html.

security concerns have been raised as the justification for potential further tightening of US export controls to bar the sale of even these downgraded chips.[47]

But further restrictions impede the development and manufacturing of chips within China: Dutch manufacturer ASML, the leading firm in manufacturing the equipment used in chipmaking, likewise faces export controls instituted by the Netherlands preventing them from maintaining, repairing, and providing spare parts for certain controlled equipment without obtaining prior government approval.[48] ASML is the primary manufacturer of extreme ultraviolet (EUV) lithography machines, used to make advanced chips, as well as immersion deep ultraviolet (DUV) lithography machines used in manufacturing memory chips.[49] These prohibitions mean that even existing hardware in China cannot be repaired by ASML when it goes out of order without the Dutch government's sign-off, significantly impacting the operations of Chinese semiconductor firms.

Compute is also an important element in many countries' national strategies on AI research and development—not just the United States'. National governments have made investments on the order of hundreds of millions of dollars to increase access to compute for academic researchers and home-grown start-ups (by contrast, Amazon recently announced a $35 billion investment in its data centers in the US state of Virginia alone).[50] At present, even these massive investments remain insufficient to compete with those of leading industry players, which are drawing on reserves orders of magnitude larger. But this push to "democratize" access to compute proceeds from the knowledge that compute-intensive research is largely dominated by industry, even in academic settings: in recent years, the largest academia-developed model used only 1 percent of the compute used to train the largest industry model.[51] Concerns over the imbalance between industry and academic research in AI is the driving premise of the National AI Research Resource (NAIRR),[52] and led the UK to announce plans to spend £100 million on acquiring compute for the country's benefit.

---

[47] Ana Swanson, David McCabe, and Michael Crowley, "Biden Administration Weighs Further Curbs on Sales of A.I. Chips to China," *New York Times,* June 28, 2023, https://nytimes.com/2023/06/28/business/economy/biden-administration-ai-chips-china.html.

[48] Cagan Koc, Jillian Deutsch, and Alberto Nardelli, "ASML Faces Tighter Dutch Restrictions on Servicing Chip Equipment in China," Bloomberg, July 14, 2023, https://bloomberg.com/news/articles/2023-07-14/asml-faces-tighter-restrictions-on-servicing-chip-gear-in-china.

[49] Arjun Kharpal, "Netherlands, Home to a Critical Chip Firm, Follows U.S. with Export Curbs on Semiconductor Tools," June 30, 2023, https://cnbc.com/2023/06/30/netherlands-follows-us-with-semiconductor-export-restrictions-.html.

[50] Matthew Barakat, "Virginia, Amazon announce $35 billion data center plan," Associated Press, https://apnews.com/article/technology-data-management-and-storage-amazoncom-inc-virginia-business-c75df1f34069b09549fe15c99335b8fb.

[51] Tamay Besiroglu et al., "The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?," forthcoming.

[52] "National AI Initiative," National Science Foundation, accessed September 20, 2023, https://nsf.gov/cise/national-ai.jsp. The proposed remedy may not address some of the fundamental concerns underlying concentration of access to computational resources and centralized industry control; the initial proposal for the NAIRR was structured as a licensing regime that would in all likelihood result in a contractual arrangement with one of the large cloud infrastructure providers. See AI Now Institute, "Democratize AI? How the Proposed National AI Research Resource Falls Short," October 5, 2021, https://ainowinstitute.org/publication/democratize-ai-how-the-proposed-national-ai-research-resource-falls-short.

All of these proposals require closer scrutiny: the original version of the NAIRR proposal would have amounted to one or a few licensing contracts with one of the Big Tech cloud infrastructure providers, and providing access to computational infrastructure will not, of its own accord, shift the market dynamics that created high levels of concentration in the first place. But what is clear is that the geopolitical terrain around computational power is an active and contested space, one in which countries are combining a variety of policy tools—state subsidies, investments and restrictive export controls—to assure the dominance of homegrown enterprises and stave off competitors.

How Much Compute Is Used in Large-Scale AI Models, and What Does It Cost?

A tremendous amount. On average, large-scale AI models use about 100 times more compute than other contemporaneous AI models.[53] If model sizes continue growing along the current trajectory, some estimates place compute costs in excess of the entire US GDP by 2037.[54] Despite this, AI models keep getting larger because size is now correlated with capability. **Competition in the market for large-scale AI models remains closely tied to the scale of the model: while factors including data quality and training method are important influences on model performance, anyone wishing to compete in the market for large-scale AI models will have to end up building larger models than the current state of the art.** Those seeking to build AI systems for particular use cases won't necessarily need to build new models from scratch—but they will be reliant on hosted models or access to APIs that, as a rule, flow through a contract with one of the major cloud infrastructure providers.

Compute costs are predictably large: the final training run of GPT-3 is estimated to have cost somewhere between $500,000 to $4.6 million.[55] Training GPT-4 may have cost in the vicinity of $50 million,[56] but the overall training cost is probably more than $100 million because compute is required for trial and error before the final training run.

[53] Sevilla et al., "Compute Trends across Three Eras of Machine Learning."

[54] This measures only the cost of the final training run for model development. See: Lennart Heim, "This Can't Go On(?) - AI Training Compute Costs," *heim.xyz* (blog), June 1, 2023, https://blog.heim.xyz/this-cant-go-on-compute-training-costs.

[55] Appenzeller, Bornstein, and Casado, "Navigating the High Cost of AI Compute."

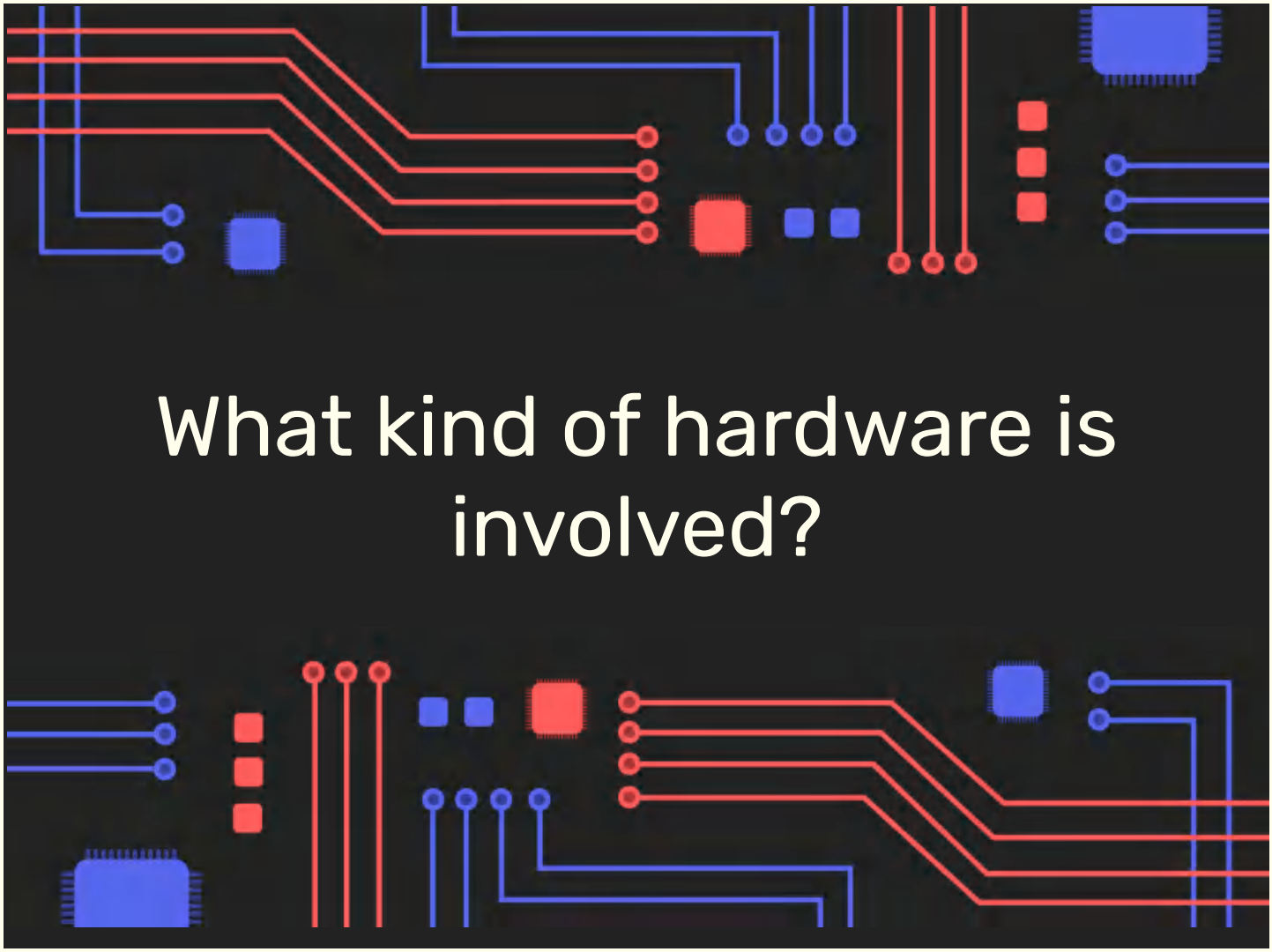
[56] Google Colab, "[PUBLIC] Cost Estimates for GPT-4," accessed July 12, 2023, https://colab.research.google.com/drive/1099z9b1l5066bT78r9ScslE_nOj5irN9?usp=sharing.

> The overall training cost of GPT-4 is probably more than $100 million because compute is required for trial and error before the final training run.

Compute is required both for training a model and for running it. For instance, one GPT-4 training run requires a huge amount of compute. But every question posed to ChatGPT also uses compute in generating a response. The latter is known as ***inference***. An inference happens anytime a model generates a response. The cost per inference is very low, but the total costs of inference can be higher than the total cost of training when hundreds of millions of inferences are made: as usage of a large-scale AI model increases, the number of inferences increases, so the compute needed to handle those inferences also increases. An OECD report cited a large cloud compute provider's estimates that its enterprise customers spent 3–4.5 percent of their total compute infrastructure expenditure on training, and 4–4.5 percent on inference.[57] Finding accurate numbers on inference costs is challenging, as companies are increasingly treating this as competitively secret information. We don't have the latest numbers for the largest AI models. Note that the numbers in this section also do not include the costs of energy and operations, which can likewise be considerable.[58]

[57] OECD, "A Blueprint for Building National Compute Capacity for Artificial Intelligence," OECD Digital Economy Papers, vol. 350, February 28, 2023, https://doi.org/10.1787/876367e3-en.

[58] Kate Saenko, "A Computer Scientist Breaks Down Generative AI's Hefty Carbon Footprint," *Scientific American*, May 25, 2023, https://scientificamerican.com/article/a-computer-scientist-breaks-down-generative-ais-hefty-carbon-footprint.



To build a large-scale AI system, the type of hardware used for training and inference is an important factor. Generally speaking, compute demand for AI has grown faster than the performance of this hardware has been able to keep up with.[59] With the success of large-scale AI, there is increasing demand for state-of-the-art chips, like Nvidia's H100 GPUs, for training AI models: using non-SOTA AI chips increases both the energy consumption and time taken for the training process by a significant amount.[60] This translates directly into cost increases: SOTA AI chips are 10–1,000 times more cost-effective than SOTA CPUs, and 33 times more cost-effective than trailing node AI chips.[61] Thus a large AI model built on trailing node AI chips would be at least 33 times more expensive than models using leading node AI chips.

[59] OECD, "A Blueprint for Building National Compute Capacity for Artificial Intelligence."

[60] Saif M. Khan and Alexander Mann, "AI Chips: What They Are and Why They Matter," Center for Security and Emerging Technology (CSET), April 2020, https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter.

[61] Ibid.

> The overall training cost of GPT-4 is probably more than $100 million because compute is required for trial and error before the final training run.

Certain types of chips are designed to be most efficient in the training and operation of large-scale AI:

- Graphics Processing Units (GPUs): GPUs were initially designed for image processing and computer graphics, and excel at running many small tasks at the same time. Because of this, they are well suited for building AI systems: they can carry out calculations in parallel while sacrificing some precision, improving AI-related efficiency over Central Processing Units (CPUs), which are designed to serve as the core component of a computer to run a centralized operating system and its applications. The design of GPUs is helpful because deep learning models can be trained using several similar calculations carried out in parallel. **Parallelism** can also be enhanced by connecting multiple AI chips together. Even the memory architecture of a GPU is optimized for AI uses. Some GPUs can store an entire large AI model on a single chip. GPUs are used primarily in the training phase, and sometimes for inference.[62]
- Field Programmable Gate Arrays (FPGAs): FGPAs can be useful to deep learning by accelerating the ingestion of data, speeding the AI workflow. These chips can offer some advantages over GPUs for certain tasks, like speech recognition and natural language processing. They are generally used for inference, but coding on FPGAs is very time consuming.[63]
- Application-Specific Integrated Circuits (ASICs): ASICs are integrated circuits that are designed for a specific application rather than for more general purpose use. Google's Tensor Processing Units (TPUs) are sometimes categorized as ASICs.[64] Generally speaking, the software layer for current ASICs is still underdeveloped.[65]

[62] Ibid.

[63] Hooker, "The Hardware Lottery."

[64] Lennart Heim, "Transformative AI and Compute," Effective Altruism Forum, September 23, 2021, https://forum.effectivealtruism.org/s/4yLbeJ33fYrwnfDev.

[65] Punch Card Investor, "Nvidia - Part 3: Beyond GPUs, Software Moat, and Competition," Substack, , September 18, 2022, https://punchcardinvestor.substack.com/p/nvidia-part-3-beyond-gpus-software.

Nvidia's H100

Nvidia's H100 GPU now sets the industry standard for computational efficiency and performance per dollar, particularly suited for training large-scale AI systems.[66] Nvidia advertises that the H100 runs 9 times faster than the A100 for training and up to 30 times faster for inference, while MosaicML reports somewhat lower figures that place the H100 running about 2.3 times faster than A100 GPUs for training and 3.5x faster for inference.[67] These systems remain prohibitively expensive for most: one 8-GPU H100 server setup runs approximately $300–400K.[68]

The companies that have gained access to H100s currently have a leading advantage in powering their large-scale AI systems, and the waiting list for them remains long. Only TSMC is currently able to fabricate H100s, contributing to the limited supply. Some observers have noted that Nvidia is making its largest allocations to smaller cloud firms, limiting its allocations to the firms that are attempting to compete with their own proprietary chips (AWS, Google and Microsoft Azure).[69] For example, Nvidia gave CoreWeave, a former crypto mining operation, early access to its H100s in November 2022, making an investment in the company soon afterward.[70]

Knowledge of who has access to H100s is itself highly competitive information. Based on the existing evidence, these include:

[66] Nvidia, "NVIDIA Announces Hopper Architecture, the Next Generation of Accelerated Computing," press release, March 22, 2022, https://nvidianews.nvidia.com/news/nvidia-announces-hopper-architecture-the-next-generation-of-accelerated-computing.

[67] MosaicML NLP Team, "MPT-30B: Raising the Bar for Open-Source Foundation Models," *Mosaic* (blog), June 22, 2023, https://mosaicml.com/blog/mpt-30b.

[68] GPU Utils, "Nvidia H100 GPUs: Supply and Demand," 1n, July 2023, https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand/#fn:1.

[69] See Sharon Goldman, "CoreWeave Came 'Out of Nowhere.' Now It's Poised to Make Billions off AI with its GPU Cloud," *Venture Beat*, accessed September 20, 2023, https://venturebeat.com/ai/coreweave-came-out-of-nowhere-now-its-poised-to-make-billions-off-of-ai-with-its-gpu-cloud/; and Anissa Gardizy, "AI Agenda: The Mysterious AI Data-Center Startup Hiring from AWS, Azure, Meta," The Information, July 27, 2023, https://theinformation.com/articles/ai-agenda-the-mysterious-ai-data-center-startup-hiring-from-aws-azure-meta.

[70] "HGX H100: The NVIDIA HGX H100 Is Here, and So Are Supercomputer Instances in the Cloud," CoreWeave, accessed September 20, 2023,

https://coreweave.com/products/hgx-h100; and Kyle Wiggers, "CoreWeave, a GPU-focused cloud compute provider, lands $221M investment," *TechCrunch*, April 20, 2023, https://techcrunch.com/2023/04/20/coreweave-a-gpu-focused-cloud-compute-provider-lands-221m-investment.

- Microsoft Azure, and select clients
    - OpenAI
    - Inflection AI
- Google Cloud
- Amazon Web Services
- Oracle
- Mistral AI
- Anthropic
- CoreWeave
- Lambda Labs
- Venture firms including C2 Investments[71]

What are the components of compute hardware?

1. Logic

The first and most central component of compute hardware is its processing power (or ***logic***). This is measured in FLOP/s, or how many computations a given resource can carry out in a second.

Peak FLOP/s refers to the computing power potential of a chip, or how many computations the chip could carry out in theory. *Effective FLOP/s* refers to a chip's real-world performance. There is always a gap between the potential effectiveness of a chip and how it performs in the real world: this is a product of practical constraints including memory, network connections or **interconnect**, and the type of **parallel processing** and software used.[72]

AI hardware is built for ***parallelism***, or the processing of many mathematical operations at the same time. In practice, algorithms never perfectly map onto the hardware and there is less than 100 percent utilization of the parallelism capacity of hardware.[73]

---

[71] Kate Clark, "Billion-Dollar AI Venture Fund Offers Elusive Nvidia Chips to Win Deals," *The Information*, June 20, 2023, https://theinformation.com/articles/former-github-ceos-novel-investment-offer-to-ai-founders-rare-server-chips.

[72] Lennart Heim, "Compute Research Questions and Metrics - Transformative AI and Compute [4/4],"Forum, *LessWrong* (blog), November 29, 2021, https://lesswrong.com/posts/G4KHuYC3pHry6yMhi/compute-research-questions-and-metrics-transformative-ai-and.

[73] Ibid.

2. Memory

Memory is a core component of compute hardware, and is where information is stored on a short-term basis after processing. Memory capacity technology is currently bottlenecked for large-scale AI:[74] Nvidia's H100 has 80 GB of memory, but the largest AI model can require many times this amount. Addressing this problem involves both practical and technological limits. An important technological constraint with current memory technology is that while logic has only one goal to optimize for (maximizing the number of transistors on a chip), memory is trying to optimize for multiple goals (**capacity, bandwidth, and latency**).[75] Latency has usually lagged behind the other two.

Unlike with logic transistors, shrinking memory cells beyond a certain point makes them less reliable, increases latency, and decreases energy efficiency.[76] In traditional CPUs, memory tends to account for over half the cost of a server setup.[77] While we know less about the memory costs in GPUs, what is clear is that they remain significant.

There are some avenues under exploration to break through this bottleneck. For example, one approach might be "compute in memory," which refers to an integration of RAM with processing elements.[78] There have been attempts to create compute in memory, but memory and compute are very difficult to manufacture together because they have differing levels of complexity and different goals. And improvements in software may hold promise: recently developed methods, like QLoRA, allow for fine-tuning an LLM on a single GPU by reducing memory requirements.[79]

In traditional CPUs, memory tends to account for over half the cost of a server setup.

[74] Asianometry, "AI's Hardware Problem," YouTube video, 16:46, December 5, 2022, https://youtu.be/5tmGKTNW8DQ.

[75] Ibid.

[76] Ibid.

[77] Dylan Patel, "CXL Enables Microsoft Azure to Cut Server Capital Expenditures by Hundreds of Millions of Dollars', *SemiAnalysis* (blog), Substack, July 7, 2022, https://www.semianalysis.com/p/cxl-enables-microsoft-azure-to-cut.

[78] This is different from SRAM, which is generally used for cache memory and not main memory due to its cost and size. Nvidia competitor Cerebras's massive wafer-scale engine, which has everything on-chip, is a type of SRAM technology.

[79] Jack Clark, "Import AI 331: 16X Smaller Language Models; Could AMD Compete with NVIDIA?; And BERT for the Dark Web," *Import AI* (blog), Substack, May 29, 2023, https://importai.substack.com/p/import-ai-331-16x-smaller-language.

3. Interconnect

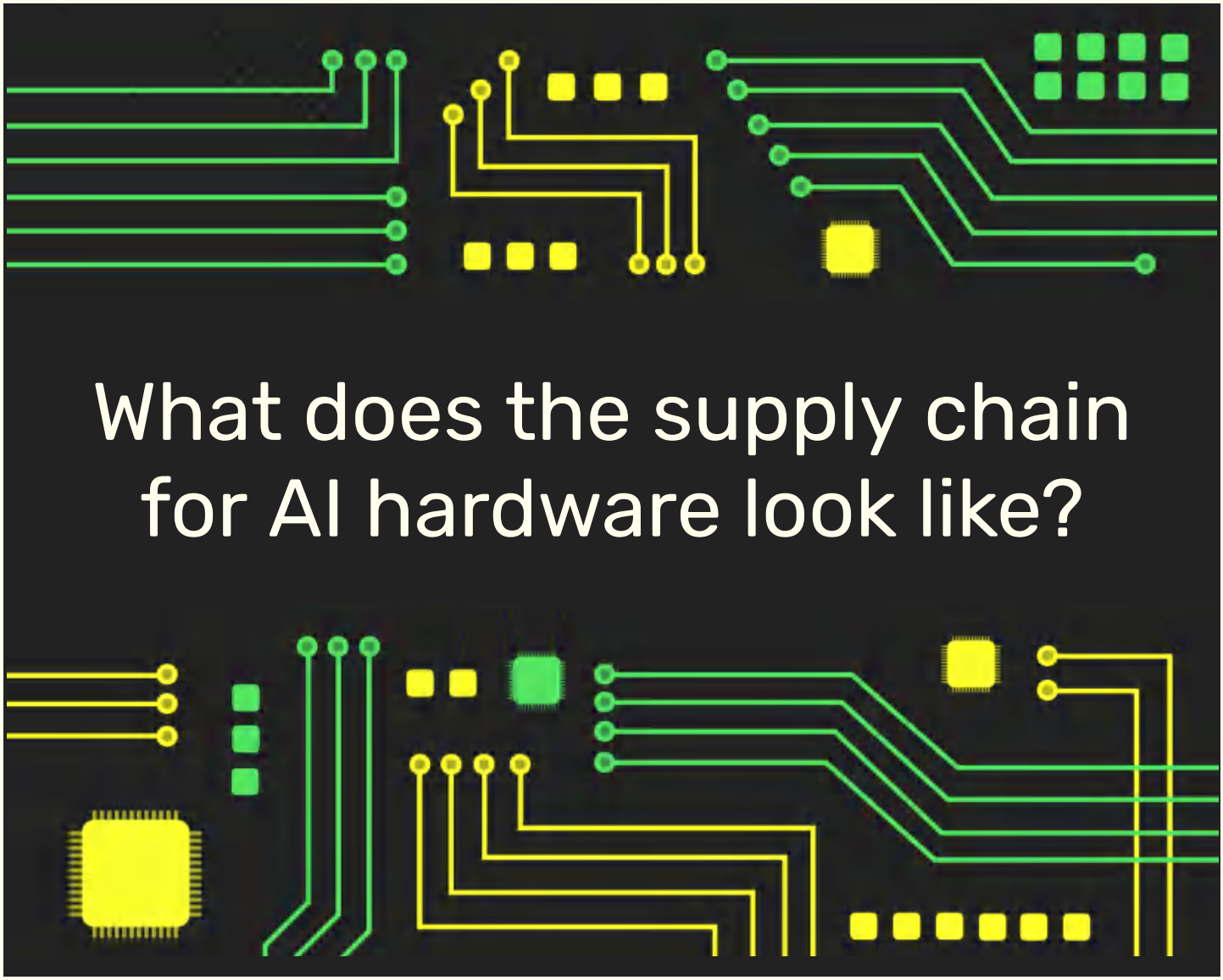
Interconnect on a chip refers to transferring information between its logic and memory components. The term can also refer to connections between chips and between servers. InfiniBand networking is the preferred standard, used in 63 of the top 100 supercomputer clusters,[80] and is used by Azure, CoreWeave, and Lambda Labs. Interconnect remains a significant component in the cost of building and operating AI.[81]

While FLOP/s grew by more than 6 times between Nvidia's A100 chips and its latest H100 chips, memory bandwidth (interconnect) only grew by 1.65 times.[82] Apart from practical and technological constraints, there is an energy cost to memory bandwidth, with a significant portion of a chip's energy usage being attributed to interconnect. Overall, this means that interconnect is an important current constraint to the growth of computational power.

[80] "About InfiniBand," InfiniBand Trade Associations, accessed September 20, 2023, https://infinibandta.org/about-infiniband.

[81] GPU Utils, "Nvidia H100 GPUs: Supply and Demand.

[82] Dylan Patel, "How Nvidia's CUDA Monopoly in Machine Learning Is Breaking - OpenAI Triton and PyTorch 2.0," *SemiAnalysis* (blog), Substack, January 16, 2023, https://www.semianalysis.com/p/nvidiaopenaitritonpytorch.



The global semiconductor market is valued at $1.4 trillion,[83] and is characterized by a series of choke points along its supply chain, aided in part by high production costs and specialized knowledge. Chips are first designed with the aid of electronic design automation software. They are then produced (or "fabricated") at a fabrication facility, after which they are assembled, tested, and packaged.

[83] "Semiconductor Market Size & Share | Industry Growth [2029], Fortune Business Insights, accessed July 12, 2023, https://fortunebusinessinsights.com/semiconductor-market-102365.

CHIP DESIGN	CHIP FABRICATION	DATA CENTERS
• Nvidia • AMD • Intel • Arm • Broadcom	• Taiwan Semiconductor Manufacturing Company • Samsung • Intel	• Google • Amazon Web Services • Microsoft • Oracle • CoreWeave • Lambda Labs

Chip Design

The design of chips for efficiently training and operating AI systems is a critical element in AI compute. There are only three designers of SOTA GPUs: Intel, AMD, and Nvidia. Of these, Nvidia has a considerable advantage: its chips are widely considered to be leading class, with the next-tier companies at least a year behind in development. Moreover, its proprietary CUDA compiling software is the most well known to AI developers, which further encourages the use of Nvidia hardware as other chips require either more extensive programming or more specialized knowledge.

Several cloud infrastructure providers are designing or plan to design their own proprietary chips: for example, **Google** designs its own chips,[84] which are currently manufactured by Samsung and soon will be manufactured by TSMC; these are called *tensor processing units* (TPUs), which it uses in addition to GPUs and CPUs. TPUs were used in the development of Google's Gemini model - an example of software and hardware integration that can impact an entire ecosystem and lead to stronger monopolization.[85] Anthropic and Midjourney use TPUs as well. Google's TPU v4 was used to train PaLM and PaLM 2 and is supposedly faster than the A100.[86] **Microsoft** announced it intends to release its own proprietary chip, called Athena, in 2024.[87] **Amazon Web Services** has its own inference and training chips.

Advancement in chip design is described in terms of ***nodes***. Node names were once used to describe the actual lengths of transistors, but as the size of leading chips has shrunk over time, they no longer do. A new node now usually means 70 percent of the transistor length

[84] Belle Lin, "In Race for AI Chips, Google DeepMind Uses AI to Design Specialized Semiconductors," *Wall Street Journal*, July 20, 2023, https://wsj.com/articles/in-race-for-ai-chips-google-deepmind-uses-ai-to-design-specialized-semiconductors-dcd78967.

[85] Dylan Patel and Daniel Nishball, "Google Gemini Eats the World – Gemini Smashes GPT-4 by 5X, the GPU-Poors," *SemiAnalysis* (blog), Substack, August 27, 2023, https://www.semianalysis.com/p/google-gemini-eats-the-world-gemini.

[86] Matthew Gooding, "Why Nvidia Won't Be Worried by Google's AI Supercomputer Breakthrough," *Tech Monitor* (blog), April 6, 2023, https://techmonitor.ai/technology/cloud/google-ai-supercomputer-nvidia-h100.

[87] Holmes and Gardizy, "Microsoft and Oracle Discussed Sharing AI Servers to Solve Shortage."

of a previous node, i.e., doubling the number of transistors per unit area, proportionately increasing the computing capacity of the chip. As of 2020 the "leading node" produced by chip designers was 5nm, and both Samsung and TSMC are producing 3nm chips with varying success.[88] The next node is expected to shrink even further to 2nm.

Nvidia's Market Dominance and Competitors

As of now, Nvidia is the undisputed market leader in SOTA AI chips and has the high margins to show for it. OpenAI used Nvidia's A100s to train the model behind ChatGPT, and now plans to use its latest H100s as part of Microsoft's Azure supercomputer.[89] Meta too has used H100s for its Grand Teton AI supercomputer, but OpenAI's ChatGPT alone is predicted to bring in a revenue of $12 billion for Nvidia over the next year. Nvidia cites its strong relationship with Taiwan Semiconductor Manufacturing Company as critical to the company's success: in a speech given at the TSMC founder's retirement celebration, Nvidia CEO Jensen Huang said that "without TSMC, there would be no Nvidia today."[90]

AMD has 20 percent of the market share in the consumer GPU market, and a chip (MI200, 6nm node) that supposedly outperforms the A100 in many metrics.[91] Its memory bandwidth is comparable to the H100. AMD has also acquired Xilinx, which makes FPGAs—which, we will recall, are very useful for inference. Even if AMD lags behind Nvidia, it might serve as an

[88] Alan Patterson, "TSMC's 3-nm Push Faces Tool Struggles," *EE Times*, April 25, 2023, https://eetimes.com/tsmcs-3-nm-push-faces-tool-struggles; Tobias Mann, "Samsung 'closing the gap' with TSMC on 3nm, 4nm," *Register*, July 18, 2023, https://theregister.com/2023/07/18/samsung_tsmc_processor_plans.

[89] Matt Vegas, "Azure Previews Powerful and Scalable Virtual Machine Series to Accelerate Generative AI,"', *Azure Blog* (blog), Microsoft, March 13, 2023, https://azure.microsoft.com/en-us/blog/azure-previews-powerful-and-scalable-virtual-machine-series-to-accelerate-generative-ai.

[90] Melody Tu, "Jen-Hsun Huang Talks about TSMC Chairman Zhang Zhongmou: A Solid Partner and a True Friend," *Nvidia* (blog), June 28, 2018, https://blogs-nvidia-com-tw.translate.goog/2018/06/28/nvidia-ceo-talk-about-friendship-with-tsmc-ceo-morris-chang; Paul Mozur and John Liu, "The Chip Titan Whose Life's Work Is at the Center of a Tech Cold War," *New York Times*, August 4, 2023, https://nytimes.com/2023/08/04/technology/the-chip-titan-whose-lifes-work-is-at-the-center-of-a-tech-cold-war.html.

[91] Punch Card Investor, "Nvidia - Part 3."

attractive alternative to Nvidia's chips in the high-demand environment we currently witness.[92]

Intel lags further behind, with its Gaudi2 chip (7nm node) developed through an acquisition supposedly performing better than the A100, but not comparable to the H100.

New would-be competitors to Nvidia are also cropping up. The **tiny corp** is a promising company that is trying to build software to enable AMD's GPUs to compete with Nvidia's chips; it then plans to build its own chips.[93] **Cerebras** differentiates itself by creating a large wafer with logic, memory, and interconnect all on-chip. This leads to a bandwidth that is 10,000 times more than the A100. However, this system costs $2–3 million as compared to $10,000 for the A100, and is only available in a set of 15. Having said that, it is likely that Cerebras is cost efficient for makers of large-scale AI models. **Graphcore** makes chips using wafer-on-wafer technology that makes them potentially slightly better than Nvidia's, but not significantly.[94]

An important way in which Nvidia distinguishes itself is through software. Its software offering, CUDA, allows programmers to use Nvidia's GPUs for general-purpose uses. CUDA serves as a software barrier to entry to Nvidia's market, because a lively developer ecosystem of memory optimization and other software libraries has been built around Nvidia's chips. Using CUDA requires a lot of expertise and knowledge about the hardware itself.[95] Unsurprisingly, experts believe that the most difficult part of US export restrictions for China to overcome is the unavailability of CUDA.[96]

But this software dominance is also slowly being challenged. OpenAI developed Triton, an open-source software solution that it claims is more efficient than CUDA. Triton can only be used on Nvidia's GPUs as of now.[97] Meta developed PyTorch and then spun off the project as an open-source initiative housed under the Linux Foundation (still financially supported by Meta),

[92] Timothy Prickett Morgan, "AMD Says AI Is the Number One Priority Right Now," *The Next Platform*, May 3, 2023, https://nextplatform.com/2023/05/03/amd-says-ai-is-the-number-one-priority-right-now.

[93] Clark, "Import AI 331."

[94] Punch Card Investor, "Nvidia - Part 3"; Tiernan Ray, "Graphcore Brings New Competition to Nvidia in Latest MLPerf AI Benchmarks," ZDNET, June 30, 2021, https://zdnet.com/article/graphcore-brings-new-competition-to-nvidia-in-latest-mlperf-ai-benchmarks.

[95] Patel, "How Nvidia's CUDA Monopoly in Machine Learning Is Breaking."

[96] Eliot Chen, "The AI Lockout," *Wire China*, March 12, 2023, https://thewirechina.com/2023/03/12/the-ai-lockout-nvidia-china.

[97] Tiernan Ray, "OpenAI Proposes Open-Source Triton Language as an Alternative to Nvidia's CUDA," ZDNET, July 28, 2021, https://zdnet.com/article/openai-proposes-triton-language-as-an-alternative-to-nvidias-cuda.

and its new version performs relatively well on Nvidia's A100.[98] The benefit of PyTorch is that it can be used across a range of hardware, but on the flip side, it is not optimized for any particular chip. Google uses its own TensorFlow framework, but this is optimized for TPUs, not for other AI chips. AMD's disadvantage in comparison to Nvidia so far has been that the former lacks a comprehensive software stack. Its ROCm stack is quite distant from CUDA's capabilities and coverage. However, this gap might be closing due to a stack developed by MosaicML, a generative AI platform.[99] This stack has done better[100] than some of Nvidia's own offerings before, and is close to achieving similar performance with AMD's chips. Some questions still remain on whether MosaicML's stack can work as well with the latest iteration of chips from AMD, and if such performance improvements translate to very large-scale models as well.[101]

Even if and when comparable offerings to Nvidia's software stack are available, it is likely that switching costs will be at least moderately high as AI teams move to new software. Companies like OpenAI that already use their own software will likely still be at a slight advantage with a more competitive market for compute software.

Other factors cement Nvidia's place as the market leader. Its hardware is leading edge at the moment: the H100 has a tensor engine that is tailor-made to speed up training and inference. As of now, it also dominates in terms of total costs of operation, because its chips can be used for the entire development process.[102] And finally, its scale allows it to reinvest in software, so it can create custom industry-specific libraries.

WHICH COMPANIES DOES NVIDIA INVEST IN?

Cloud infrastructure firms:

- CoreWeave[103]

[98] Patel, "How Nvidia's CUDA Monopoly in Machine Learning Is Breaking."

[99] Dylan Patel and Aleksandar Kostovic, "AMD AI Software Solved – MI300X Pricing, Performance, PyTorch 2.0, Flash Attention, OpenAI Triton," *SemiAnalysis* (blog), Substack, June 30, 2023, https://www.semianalysis.com/p/amd-ai-software-solved-mi300x-pricing.

[100] "Better" here is defined as getting a chip to perform closer to its potential.

[101] Patel and Kostovic, "AMD AI Software Solved."

[102] Punch Card Investor, "Nvidia - Part 3."

[103] Kyle Wiggers, "CoreWeave, a GPU-Focused Cloud Compute Provider, Lands $221M Investment," *TechCrunch*, April 20, 2023, https://techcrunch.com/2023/04/20/coreweave-a-gpu-focused-cloud-compute-provider-lands-221m-investment.

Large-Scale AI firms:
- Inflection AI[104]
- Cohere[105]

Generative AI firms:
- Inworld AI (generative game development)[106]
- Runway (generative AI-based creative tools)[107]

Other AI companies:
- OmniML (model efficiency)
- Recursion (biotech)
- Skydio (AI-enabled drones)[108]
- Adept AI (language models)
- Foretellix (automated driving)[109]
- Serve Robotics[110] (sidewalk delivery robotics)
- Outrider[111] (autonomous trucks)
- Deepgram[112] (speech recognition)

Potential Future Investments:
- Lambda Labs[113] (cloud infrastructure)
- Arm[114] (chip design firm)

[104] Alex Konrad, "Inflection AI, the Year-Old Startup behind Chatbot Pi, Raises $1.3 Billion," Forbes, June 29, 2023, https://forbes.com/sites/alexkonrad/2023/06/29/inflection-ai-raises-1-billion-for-chatbot-pi.

[105] Sharon Goldman, "OpenAI rival Cohere raises a fresh $270 million to bring generative AI to the enterprise," *VentureBeat*, June 8, 2023, https://venturebeat.com/ai/openai-rival-cohere-raises-a-fresh-270-million-to-bring-generative-ai-to-the-enterprise.

[106] Jacob Robbins, "Inworld AI Brings Generative AI to Game Development Tools at $500M+ Valuation," PitchBook, August 2, 2023, https://pitchbook.com/news/articles/generative-ai-gaming-inworld-venture-funding.

[107] "Global VC Deals for July 3, 2023," PitchBook, July 3, 2023, https://pitchbook.com/newsletter/global-vc-deals-for-july-3-2023.

[108] "Global VC Deals for February 28, 2023," PitchBook, February 28, 2023, https://pitchbook.com/newsletter/global-vc-deals-for-february-28-2023.

[109] "Global VC Deals for May 3, 2023," PitchBook, May 3, 2023, https://pitchbook.com/newsletter/global-vc-deals-for-may-3-2023.

[110] "Nvidia Invests USD 10 Million in Serve Robotics," Geospatial World, October 3, 2022, https://geospatialworld.net/news/nvidia-invests-usd-10-million-in-serve-robotics.

[111] "Outrider Hauls in $73M for Autonomous Trucks," PitchBook, January 20, 2023, https://pitchbook.com/newsletter/outrider-hauls-in-73m-for-autonomous-trucks.

[112] "Deepgram Pulls in $12M," PitchBook, March 19, 2020, https://pitchbook.com/newsletter/deepgram-pulls-in-12m.

[113] Maria Heeter, Kate Clark, and Stephanie Palazzolo, "Nvidia Accelerates AI Startup Investments, Nears Deal with Cloud Provider Lambda Labs," *The Information*, July 18, 2023, https://theinformation.com/articles/nvidia-accelerates-ai-startup-investments-nears-deal-with-cloud-provider-lambda-labs.

[114] CPI, "Nvidia Considers Investing in Arm's IPO," Competition Policy International, July 12, 2023, https://competitionpolicyinternational.com/nvidia-considers-investing-in-arms-ipo.

Chip Fabrication

The entry cost of the chip fabrication market is astronomical, and concentration in the semiconductor industry increases as transistor size decreases (i.e., as chip capability increases). The three chipmakers for leading node chip production are TSMC (makes roughly 70–80 percent of the revenue), Samsung, and Intel (currently both of these fabricators are about a year behind TSMC). In 2019, the cost of production of a next-generation chip was estimated to be around $30–80 million.[115] In 2017, the fixed cost of building a fabrication facility was estimated to be around $7 billion.[116] Today, that cost is more than $20 billion.[117]

TSMC is the only company currently manufacturing both Nvidia and AMD's high-end chips, and is the only company making Intel's Arc GPUs (though Nvidia appears to be exploring other options).[118] It is the sole chip manufacturer capable of producing leading 3nm and 2nm nodes, and as such has the capacity to direct whose chips are prioritized, and who gains access to their leading manufacturing processes. Any potential competitor to Nvidia in chip design (as Google, Microsoft, and AWS are all attempting to be) will have to navigate the TSMC's stranglehold on chip manufacturing at scale.

Fabrication costs continue to grow at 11 percent per year and chip design costs are growing at 24 percent per year, while the semiconductor market grows at only 7 percent per year.[119] High fixed costs mean high barriers to entry, but **soon costs may be high enough that even a natural monopoly cannot recoup them.**[120] There is also a monopoly on obtaining the equipment needed to manufacture leading node chips: Dutch manufacturing firm ASML is the only firm capable of producing the photolithography equipment required for the leading node.

Chip Assembly, Testing, and Packaging (ATP)

ATP involves cutting wafers into chips and adding wire connectors to chip frames. It can occur in-house or be outsourced. ATP work has generally been outsourced to developing

[115] Michael Feldman, "The Era of General Purpose Computers Is Ending," *The Next Platform*, February 5, 2019, https://nextplatform.com/2019/02/05/the-era-of-general-purpose-computers-is-ending.

[116] Hooker, "The Hardware Lottery."

[117] Veedrac, "Moore's Law, AI, and the Pace of Progress," Forum, *LessWrong* (blog), December 11, 2021, https://lesswrong.com/posts/aNAFrGbzXddQBMDqh/moore-s-law-ai-and-the-pace-of-progress.

[118] Paul Alcorn, "Nvidia CEO Says Intel's Test Chip Results For Next-Gen Process Are Good," *Tom's Hardware*, May 30, 2023, https://tomshardware.com/news/nvidia-ceo-intel-test-chip-results-for-next-gen-process-look-good.

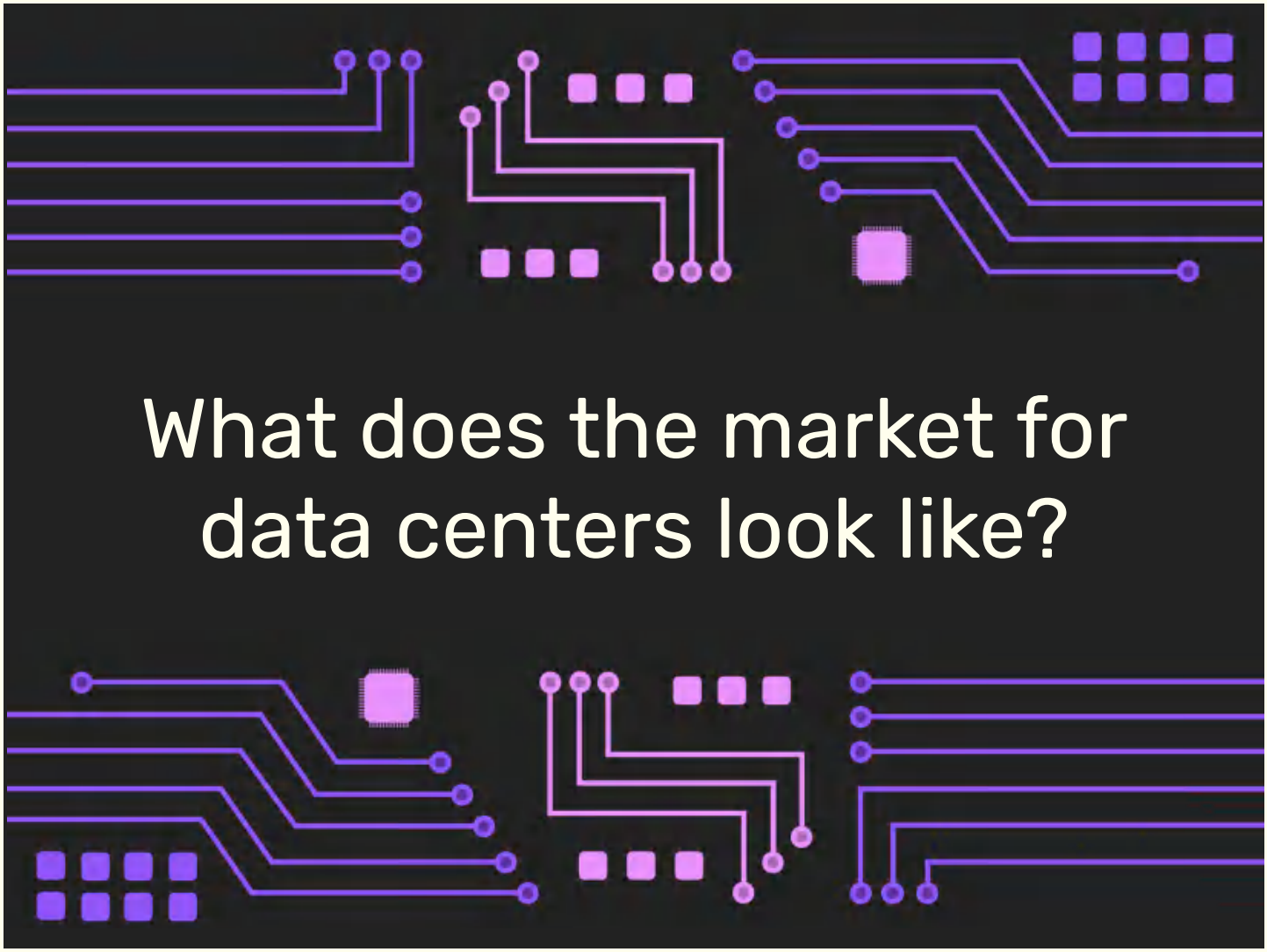
[119] Khan and Mann, "AI Chips."

[120] Ibid.

countries, including China. China is particularly competitive in integrated assembly.[121] Of particular note is that although TSMC is building a chip fabrication facility in Arizona under the US CHIPS Act, all of the chips fabricated at this facility will need to be sent to Taiwan for packaging, meaning that TMSC's US chip production process will still require a global supply chain network.[122]

[121] Saif M. Khan, Dahlia Peterson, and Alexander Mann, "The Semiconductor Supply Chain," Center for Security and Emerging Technology (CSET), January 2021, https://cset.georgetown.edu/publication/the-semiconductor-supply-chain.

[122] Wayne Ma, "The Flaw in Apple's Plan to Make Chips in Arizona," *The Information*, September 11, 2023, https://theinformation.com/articles/apples-plan-to-make-chips-in-arizona-tsmc-nvidia-amd-tesla.



Usually, large AI models are trained using a cluster of many chips known as an *AI supercomputer*.

Supercomputers are hosted in data centers, which provide the infrastructure to keep the hardware running. Data centers as a whole—including the connected chips within them—represent the infrastructure layer of AI compute. There are probably between 10,000 and 30,000 data centers in the world, and only about 325–1400 of these could host an AI supercomputer.[123] Large data centers enable cloud computing, a market that is well known to exhibit a high degree of concentration. See below:

[123] Konstantin Pilz and Lennart Heim, *Compute at Scale*, Konstantin F. Pilz, July 2023, https://konstantinpilz.com/data-centers/report.

COMPANY	CLOUD COMPUTING MARKET SHARE (Q1 OF 2023)
Amazon Web Services	**32 percent**
Microsoft Azure	**23 percent**
Google Cloud	**10 percent**

Adapted from source[124]

AI models' immense compute requirements affect the concentration in the market for data centers and cloud services. Cloud service providers provide compute at deep discounts to large AI research labs so as to be able to increase their own market share.[125] These agreements are made at early stages so that they can involve equity investment, sometimes being finalized even before a product is launched.[126] See examples below:

- **Microsoft has invested in OpenAI, and Azure is the exclusive cloud provider for OpenAI for both training and inference.[127] Microsoft's first AI supercomputers were built exclusively for OpenAI.[128]** It also has exclusive rights to sell OpenAI model access to its cloud customers under the Azure OpenAI Service, which is a set of ready-to-use AI APIs.[129] The following lines from OpenAI's blog gesture toward the degree of cooperation between the two: *"In an effort to build and deploy safe AI systems, our teams regularly collaborate to review and synthesize shared lessons—and use them to inform iterative updates to our systems, future research, and best practices for use of these powerful AI systems across the industry."*[130]
- **Google Research's Brain team and DeepMind have now been fully integrated into Google DeepMind.** The two teams' competition for data center time was the most

[124] Mark Haranas, "AWS, Microsoft, Google's Cloud Market Share Q1 2023," CRN, May 4, 2023, https://crn.com/news/cloud/aws-microsoft-google-s-cloud-market-share-q1-2023.

[125] Pilz and Heim, *Compute at Scale*.

[126] Murgia, "Big Tech Companies Use Cloud Computing Arms to Pursue Alliances with AI Groups."

[127] OpenAI, "OpenAI and Microsoft Extend Partnership," press release, January 23, 2023, https://openai.com/blog/openai-and-microsoft-extend-partnership.

[128] Eric Boyd, "General Availability of Azure OpenAI Service Expands Access to Large, Advanced AI Models with Added Enterprise Benefits,", *Azure Blog* (blog), Microsoft, January 17, 2023, https://azure.microsoft.com/en-us/blog/general-availability-of-azure-openai-service-expands-access-to-large-advanced-ai-models-with-added-enterprise-benefits.

[129] Holmes and Gardizy, "Microsoft and Oracle Discussed Sharing AI Servers to Solve Shortage."

[130] OpenAI, "OpenAI and Microsoft Extend Partnership."

critical factor driving the integration.[131] It's worth noting that both OpenAI and DeepMind made decisions to be invested in and acquired,[132] respectively, primarily due to the costs of compute.

- **Google Cloud is Anthropic's 'preferred' cloud partner-[133] or was, until Amazon took a minority stake worth up to $4B in the company.[134]** Google has also invested $300 million in Anthropic for a 10 percent stake. Google Cloud is also the preferred[135] cloud partner for Cohere, which is in talks with Nvidia and others for funding.[136]

- **Amazon has entered into partnerships with open source model developers and platforms.** For instance, Hugging Face has partnered with AWS in a revenue-sharing agreement to allow developers using Hugging Face to use AWS's compute and software.[137] Amazon has brought its partnerships together to create Amazon Bedrock, an API service that gives access to models from AI21 Labs, Anthropic, Stability AI, and AWS itself, replicating the 'app store' or 'marketplace' model that has elsewhere raised concerns about self-preferencing and rent-seeking behavior.

- **Oracle is a surprise success story in this segment, even if it currently holds only 5 percent of the cloud market.** It offers compute credits worth several hundred thousand dollars to AI startups.[138] Notable customers include Character.ai and Adept AI Labs. Oracle is able to provide compute that is cheaper and faster than its competitors' because it has optimized its cables connecting GPUs for machine learning, and has been optimizing its hardware for machine learning for a while.[139] Some AI startups choose Oracle because it does not directly compete with them, but Oracle has reportedly considered developing its own LLM.[140]

---

[131] Jeremy Kahn, "The Google Brain-DeepMind Merger Is Probably Good for Google. It Might Not Be for Us," *Fortune*, April 28, 2023, https://fortune.com/2023/04/28/the-google-brain-deepmind-merger-alphabet-pichai-risks-eye-on-a-i.

[132] Ibid.

[133] Murgia, "Big Tech Companies Use Cloud Computing Arms to Pursue Alliances with AI Groups."

[134] Adam Satariano and Cade Metz, "Amazon Takes a Big Stake in the A.I. Startup Anthropic", *New York Times,* September 25, 2023, https://www.nytimes.com/2023/09/25/technology/amazon-anthropic-ai-deal.html.

[135] Kevin Ichhpurani, 'Building the Most Open and Innovative AI Ecosystem', *Google Cloud Blog* (blog), 15 March 2023, https://cloud.google.com/blog/products/ai-machine-learning/building-an-open-generative-ai-partner-ecosystem.

[136] Cade Metz, 'Generative A.I. Start-Up Cohere Valued at About $2 Billion in Funding Round', *The New York Times*, 2 May 2023, sec. Technology, https://nytimes.com/2023/05/02/technology/generative-ai-start-up-cohere-funding.html.

[137] Kevin McLaughlin and Anissa Gardizy, 'After Years of Resistance, AWS Opens Checkbook for Open-Source Providers', *The Information*, 24 May 2023, https://theinformation.com/articles/after-years-of-resistance-aws-opens-checkbook-for-open-source-providers

[138] Aaron Holmes, 'AI Startups Find an Unlikely Friend: Oracle', *The Information*, 22 February 2023, https://theinformation.com/articles/ai-startups-find-an-unlikely-friend-oracle.

[139] Ibid.

[140] Holmes and Gardizy, "Microsoft and Oracle Discussed Sharing AI Servers to Solve Shortage."

- **Cloud companies also invest in AI service companies or companies that do not necessarily develop their own foundation models.** For instance, Google has invested in Runway (powered by Stable Diffusion), continuing an established practice of investing in startups to turn them into cloud customers.[141] Before the investment, Runway's preferred cloud provider was AWS.

We have seen that AI startups' choice of cloud provider is often guided by concerns about being outcompeted by cloud providers offering their own AI models.[142]

This is why some AI companies have also been hesitant to use chips designed by cloud providers, as there is a risk of being locked into a specific ecosystem.[143]

The market power of cloud providers has been of concern not only to startups, but also even to players that dominate upstream markets, such as Nvidia. Nvidia has made attempts to introduce competition to cloud markets to reduce its own costs and also to reduce the likelihood that cloud providers eat into the chip design market. It has chosen to give preferential access to H100s to smaller actors like CoreWeave and Lambda Labs instead of the largest cloud players, and has made a $100 million investment in CoreWeave that the company was then able to collateralize to raise a further $2.3 billion in debt.[144] It is in talks to make a similar investment deal with Lambda Labs.[145] Large cloud players and AI companies are thus compelled to rent compute from smaller providers preferred by Nvidia: for example, Microsoft signed a deal potentially worth billions for access to CoreWeave's GPUs.[146]

Nvidia is also muscling its way into the cloud business directly: leveraging its dominant hold on the design of state of the art chips, it used the promise of its H100 chip to make deals with Microsoft, Google, and Oracle - but notably not AWS - to lease servers in these firms' data centers so it rent them to AI software developers at a premium through a program it called DGX Cloud. The service also offers pre-trained models including Nvidia's Megatron

[141] Kate Clark et al., "Google Invests in AI Startup Runway to Wrest Cloud Business From AWS," *The Information*, May 31, 2023, https://theinformation.com/articles/google-invests-in-ai-startup-runway-to-wrest-cloud-business-from-aws.

[142] Holmes, "AI Startups Find an Unlikely Friend."

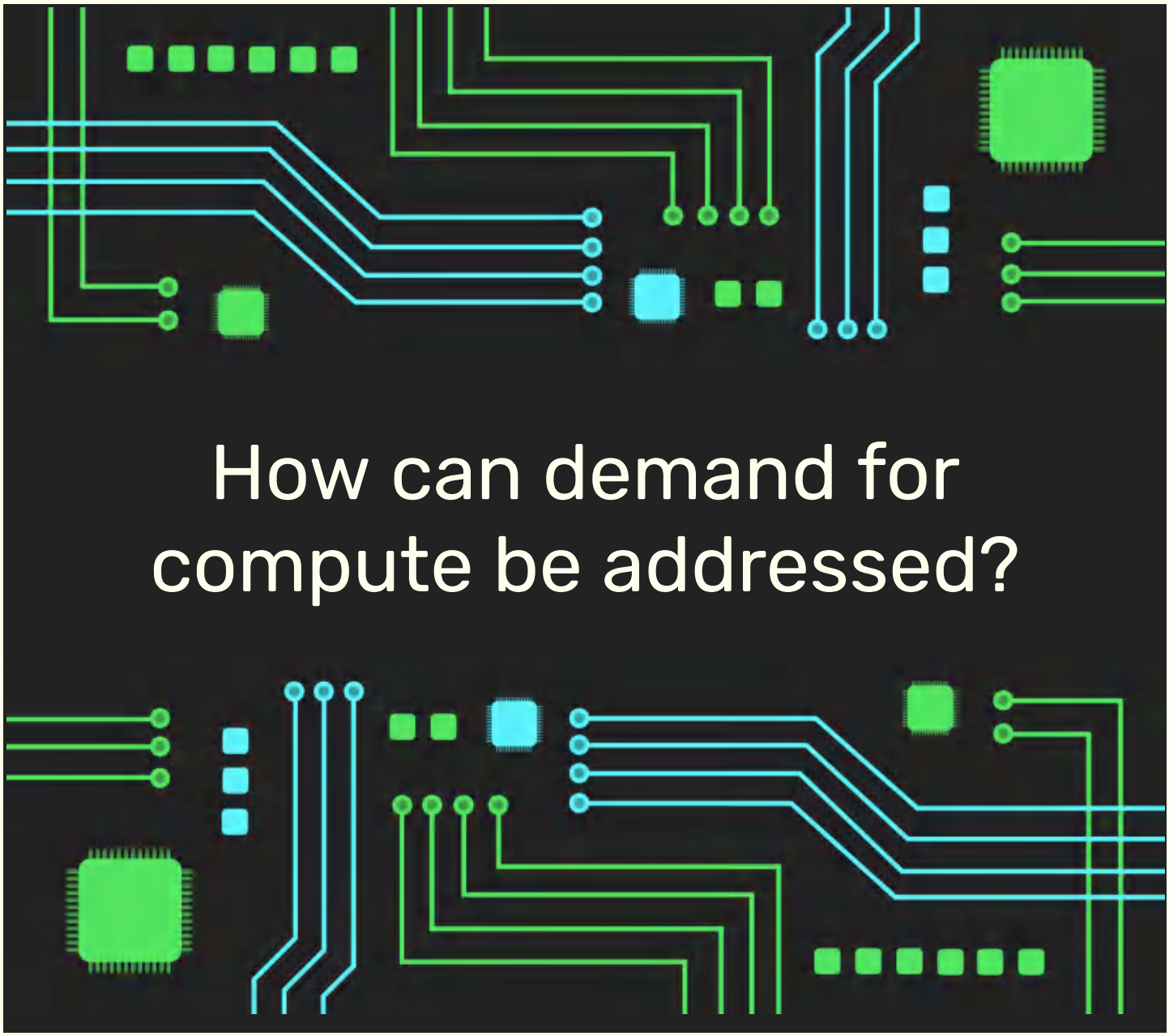
[143] Holmes.

[144] Hu, "CoreWeave Raises $2.3 Billion in Debt Collateralized by Nvidia Chips."

[145] Heeter, Clark, and Palazzolo, "Nvidia Accelerates AI Startup Investments, Nears Deal with Cloud Provider Lambda Labs."

[146] Jordan Novet, "Microsoft Signs Deal for A.I. Computing Power with Nvidia-Backed CoreWeave That Could Be Worth Billions," CNBC, June 1, 2023, https://cnbc.com/2023/06/01/microsoft-inks-deal-with-coreweave-to-meet-openai-cloud-demand.html.

530B large language model and PeopleNet, a model for recognizing humans in video footage.[147]



We are now in a position to explore the different strategies that different stakeholders can adopt to grapple with the concentrated market for AI compute.

[147] Anissa Gardizy and Aaron Holmes, "Nvidia Muscles Into Cloud Services, Rankling AWS," The Information, Sept. 11, 2023, https://www.theinformation.com/articles/nvidia-muscles-into-cloud-services-rankling-aws.

Reducing Compute Costs

One way to reduce compute costs is to make hardware improvements such that more computations can be done using fewer chips. While chips have advanced considerably over the past few years, experts differ on whether this pace is slowing down.

Don't Count on Moore's Law

The most well-known benchmark for chip performance progress is *Moore's law*, which predicts that the number of transistors on chips doubles roughly every two years (more transistors implies faster computations). Some experts believe that Moore's law will inevitably slow down as transistors approach physical limits of size—they are now only a few atoms wide.[148] Additionally, the rate of speed and efficiency improvements from an increase in the number of transistors is itself slowing but consistent.[149]

Others believe that innovations always emerge to keep growth consistent at a Moore's law level.[150] They use DNA as an indication that physical and chemical switches can be much much smaller and that it is physically possible for very small items to carry a lot of information. They point out that "physical limits" does not mean limits set by electromagnetic signals, silicon, and limited investment, all of which are ultimately variable.[151]

Overall, it is likely that some discontinuities or innovations are required to continue Moore's law-level improvements, but one can argue that these have always been required.

Another way to reduce compute costs is to use existing compute more efficiently. This can be done by making smarter algorithms that use less compute for the same output. In 2020, OpenAI estimated that training a 2012 model required 44 times less compute in 2020 than

[148] Khan and Mann, "AI Chips."

[149] Ibid.

[150] Veedrac, "Moore's Law, AI, and the Pace of Progress."

[151] Ibid.

it did originally in 2012.[152] By comparison, if there was no algorithmic improvement and only a Moore's law-level improvement in hardware during this period, we would need 11 times less compute to train a 2012 model in 2020.[153] This means that algorithmic progress

[152] Danny Hernandez and Tom B. Brown, "Measuring the Algorithmic Efficiency of Neural Networks," arXiv, May 8, 2020, https://doi.org/10.48550/arXiv.2005.04305.

[153] Ibid.

(compute efficiency) has contributed more toward performance improvement than pure hardware efficiency. Notably, Nvidia has been making strategic acquisitions of companies focused on exactly this.[154]

Algorithmic efficiency is already a domain along which the largest AI players compete fiercely. However, much more efficiency is theoretically still possible even with current compute capacity.[155]

Using Smaller Models

AI researchers could reduce compute use by abandoning the current method of building very large models and instead seeking capability improvements through smaller models. There are few indicators that companies are seriously pursuing this strategy, even despite considerable harms associated with large-scale AI.[156]

At present, large-scale AI models achieve higher utilization rates, because in AI training, memory capacity and bandwidth demands do not grow as fast as logic demands do. Memory capacity and bandwidth are constraints on leading node chips (they improve at slower rates than logic components), so it is beneficial to scale up if logic demands grow faster than the other two.

For some time, it appeared as if training a smaller model on the outputs of a larger model could allow the smaller model to approach the capabilities of the larger model. Researchers at Stanford first demonstrated this with Alpaca, a fine-tuned version of Meta's LLaMA. Alpaca was a smaller model trained on ChatGPT outputs and appeared to develop capabilities quite close to ChatGPT's own. However, later work has shown that in such "imitation" models[157]

1. increasing the amount of imitation data does not close any capabilities gaps, while increasing the base model size does;
2. a lot of capability comes from good and extensive training data, as well as from sheer model size; and

[154] Anissa Gardizy, "Nvidia Acquired AI Startup That Shrinks Machine-Learning Models," *The Information*, June 30, 2023, https://theinformation.com/articles/nvidia-acquired-ai-startup-that-shrinks-machine-learning-models

[155] Thompson et al., "The Computational Limits of Deep Learning."

[156] AI Now Institute, *ChatGPT and More: Large Scale AI Models Entrench Big Tech Power*, April 11, 2023, https://ainowinstitute.org/publication/large-scale-ai-models.

[157] Arnav Gudibande et al., "The False Promise of Imitating Proprietary LLMs," arXiv, May 25, 2023, http://arxiv.org/abs/2305.15717.

3. it is easier to imitate a specific rather than general capability of a larger model in this way.

Overall, it now seems like imitation models are able to better capture the style of a larger model, but not its substantive capacity.[158]

Paradigm Shifts and Breakthroughs

Paradigm shifts in compute development, such as neuromorphic computing or quantum computing, could create an entirely new market structure and much higher compute capacity. However, we have yet to see these paradigms truly emerge as capable, let alone as scalable or commercially viable—and if they did emerge, the likelihood is that they would be the product of investments by the same firms already dominant in compute.

Breakthroughs in memory technology could also change the market structure for compute. We have seen before that memory is a significant bottleneck to compute capacity because it has tended to grow slower than logic. Again, we have yet to see promising breakthroughs in memory or memory bandwidth.

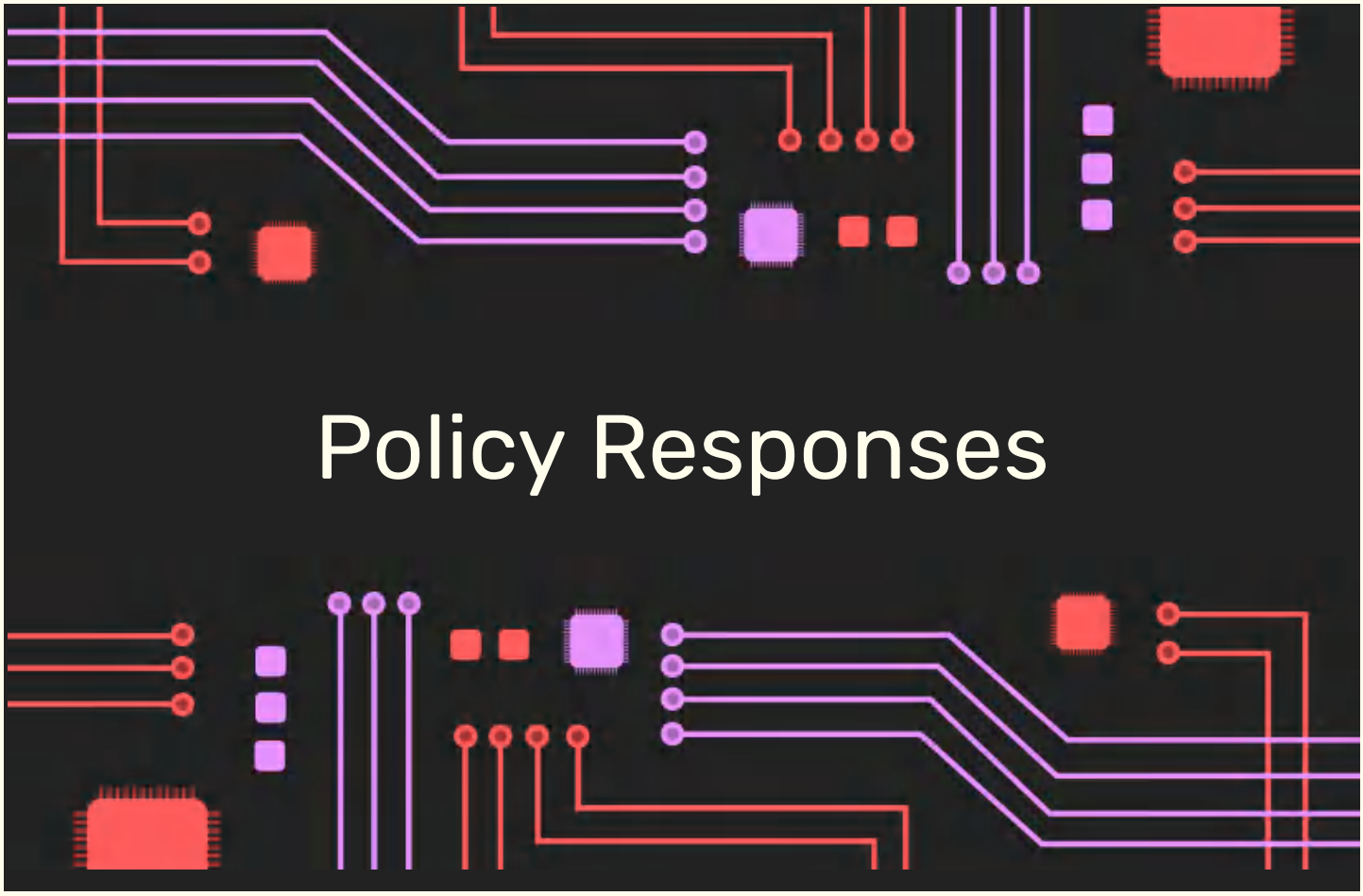
There are a few other pathways toward overcoming the compute barrier, significant among which is a degree of decentralized training. Federated learning could possibly be a way to achieve scale without centralization.[159] Federated learning works by training on data on the edge, and only transferring encrypted training results to the center. Major challenges to adopting federated learning include the following:[160]

- Data on different devices is heterogenous.
- Secure aggregation of training results is difficult.
- On-device memory requirements are high.
- There is an associated communication cost.

[158] Ibid.

[159] Neeraj Hablani, "Federated Learning at the Edge May Out-Compete the Cloud on Privacy, Speed and Cost," *VentureBeat*, February 26, 2023, https://venturebeat.com/ai/federated-learning-at-the-edge-may-out-compete-the-cloud-on-privacy-speed-and-cost.

[160] See Zain Hasan, "Running Large Language Models Privately - PrivateGPT and Beyond," *Weaviate* (blog), May 30, 2023, https://weaviate.io/blog/private-llm; and Tien-Ju Yang et al., "Online Model Compression for Federated Learning with Large Models," arXiv, May 6, 2022, https://doi.org/10.48550/arXiv.2205.03494.



Compute Power is the Emerging Frame for AI Industrial Policy

Compute power is a key facet of the emerging industrial policy frame in AI. Nations seeking a competitive advantage in AI are investing heavily in semiconductor development and undercutting their adversaries through strict export control regimes that seek to limit access across the compute supply chain on national security grounds.

United States

The United States CHIPS and Science Act of 2022 was the first major industrial policy measure passed in tech in recent history, focused on growing a national US-based semiconductor fabrication industry. Prior to the passage of the act, the US produced about 10 percent of the world's supply of semiconductors. The new Act includes measures such as these:

- $52.7 billion for American semiconductor research, development, manufacturing and workforce development
- A 25 percent investment tax credit for semiconductor manufacturing and related equipment
- $10 billion to invest in regional innovation and technology hubs
- Increased support and investment in STEM education and training, particularly supporting HBCUs and other minority-serving institutions [161]

Over 50 new semiconductor projects were announced worth $200 billion following the passage of the Act, according to the Semiconductor Industry Association.[162] Among them is TSMC, which plans to make a $40 billion investment in a new facility in Phoenix, Arizona.[163]

This is particularly notable because it illustrates that market subsidies can function to exacerbate rather than ameliorate market concentration if not carefully calibrated: given the existing bottlenecks in chip fabrication, such investments can easily be captured by dominant players even if they introduce more geographical spread.[164] Notably, the chips produced within TSMC's new facility will still be sent back to Taiwan for packaging and assembly, subverting the hope of creating fully made-in-the-USA chips.[165]

The National AI Research Resource is another compute-related policy proposal. The National AI Initiative Act of 2020 charged the National Science Foundation with developing a proposal for a shared research infrastructure that would expand access to computational power, high-quality data, and other support to facilitate more widespread AI

[161] White House, "FACT SHEET: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China," August 9, 2022, https://whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china.

[162] Robert Casanova, "The CHIPS Act Has Already Sparked $200 Billion in Private Investments for U.S. Semiconductor Production," *Semiconductor Industry Association* (blog), December 14, 2022, https://semiconductors.org/the-chips-act-has-already-sparked-200-billion-in-private-investments-for-u-s-semiconductor-production.

[163] Rishi Iyengar, "Who Will Make the Chips?," *Foreign Policy*, May 30, 2023, https://foreignpolicy.com/2023/05/30/semiconductor-chips-competition-united-states-china-engineers.

[164] Thanks to Max von Thun for this point.

[165] Wayne Ma, "The Flaw in Apple's Plan to Make Chips in Arizona," The Information, Sept. 11, 2023, https://www.theinformation.com/articles/apples-plan-to-make-chips-in-arizona-tsmc-nvidia-amd-tesla.

development.[166] The final NAIRR report articulates that a minimum of 18 providers should be part of the NAIRR to ensure resilience and prevent capture. However, under the current conditions of compute scarcity, the design of the NAIRR allows for—and would likely necessitate—some kind of licensing contract with one of the large cloud infrastructure providers, leading to critiques that the NAIRR falls short of its promise to "democratize" AI development processes.[167] The NAIRR is likely to come up for a Congressional vote in fall 2023 through the CREATE AI Act - which currently does not mandate interoperability to ensure the NAIRR can run across multiple cloud providers.[168]

France

Across the Atlantic, compute power has been an important element in France's national interest in building out its AI capabilities. Among other moves, France funded the creation of the Jean Zay supercomputer in 2019, operated by the Centre national de la recherche scientifique (CNRS). The computer was used to train BigScience's BLOOM large-scale AI model.[169] There has also been a concerted effort by non-governmental actors to prop up a homegrown startup industry in France; for example Xavier Niel recently announced a 200 million euro investment to provide compute power through a partnership with Nvidia for French AI research.[170]

United Kingdom

The United Kingdom recently announced its own plans to spend £900 million to build a supercomputer in support of UK-based AI research, and is investing £2.5 billion in quantum technologies.[171] It will also spend £100 million in direct purchases of GPUs from Nvidia, AMD, and Intel.[172]

---

[166] National Artificial Intelligence Research Resource Task Force, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource*, January 2023, https://ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf.

[167] AI Now Institute, "Democratize AI? How the Proposed National AI Research Resource Falls Short."

[168] Congresswoman Anna Eshoo, "CREATE AI Act", July 28, 2023, https://eshoo.house.gov/media/press-releases/ai-caucus-leaders-introduce-bipartisan-bill-expand-access-ai-research

[169] "Jean Zay: Introduction," Institute for Development and Resources in Intensive Scientific Computing, CNRS, accessed September 20, 2023, http://www.idris.fr/eng/jean-zay/jean-zay-presentation-eng.html.

[170] Benoit Berthelot, "Billionaire Xavier Niel Invests €200 Million in French AI Push", Bloomberg, Sept. 26, 2023, https://www.bloomberg.com/news/articles/2023-09-26/billionaire-xavier-niel-invests-200-million-in-french-ai-push

[171] Dan Milmo and Alex Hern, "UK to Invest £900m in Supercomputer in Bid to Build Own 'BritGPT'," *Guardian*, March 15, 2023, https://theguardian.com/technology/2023/mar/15/uk-to-invest-900m-in-supercomputer-in-bid-to-build-own-britgpt.

[172] Anna Isaac, "UK to Spend £100m in Global Race to Produce AI chips," *Guardian*, August 20, 2023, https://theguardian.com/business/2023/aug/20/uk-global-race-produce-ai-chips.

Japan

Japan's JFTC signaled an interest in making similar moves in its cloud market study, pointing to the increasing dominance of US-based cloud infrastructure companies as a reason for increased investment in homegrown alternatives.[173]

Enforcement Agencies Have Converged on Cloud Concentration as a Significant Problem

A handful of enforcement agencies around the globe have initiated or recently concluded studies of the cloud computing market, exploring whether there is concentration in cloud computing and its downstream effects on the economy.

The US Federal Trade Commission recently concluded an inquiry into cloud computing that particularly expressed interest in the effects of potential concentration in cloud computing on artificial intelligence.[174] It also noted the potentially harmful impact of dependencies within a large part of the economy on one or a handful of cloud providers. This concern about "single points of failure" was echoed in a Treasury Department report that worried there could be a "cascading impact across the broader financial sector."[175] Soon after the completion of the RFI, the FTC released a blog post identifying the risk of anticompetitive practices among the firms responsible for providing computational resources: "incumbents that offer both compute services and generative AI products—through exclusive cloud partnerships, for instance—might use their power in the compute services sector to stifle competition in generative AI by giving discriminatory treatment to themselves and their partners over new entrants."[176]

Particularly of note is the FTC's 2022 intervention to prevent the proposed acquisition of Arm Ltd. by Nvidia, what would have been the largest semiconductor chip merger.[177] Among

[173] Japan Fair Trade Commission, *Report Regarding Cloud Services*, June 28, 2022, https://www.jftc.go.jp/en/pressreleases/yearly-2022/June/220628.html.

[174] The FTC Office of Technology, "An Inquiry into Cloud Computing Business Practices: The Federal Trade Commission Is Seeking Public Comments," *Federal Trade Commission Technology Blog* (blog), March 22, 2023, https://ftc.gov/policy/advocacy-research/tech-at-ftc/2023/03/inquiry-cloud-computing-business-practices-federal-trade-commission-seeking-public-comments.

[175] U.S. Department of the Treasury, "New Treasury Report Assesses Opportunities, Challenges Facing Financial Sector Cloud-Based Technology Adoption," press release, February 8, 2023, https://home.treasury.gov/news/press-releases/jy1252.

[176] Staff in the Bureau of Competition & Office of Technology, "Generative AI Raises Competition Concerns," *Federal Trade Commission Technology Blog* (blog), June 29, 2023, https://ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.

[177] See Federal Trade Commission, "Nvidia/Arm, In the Matter of," February 14, 2022, https://ftc.gov/legal-library/browse/cases-proceedings/2110015-nvidiaarm-matter; the European Commission and the UK's Competition and Markets Authority also launched investigations, and the CMA was likely to block the merger before it was called off.

other justifications for the FTC's intervention was a concern that it would be harmful to innovation, inhibiting Arm from development of on-chip AI functions not tied to Nvidia's proprietary hardware. A second concern was that Nvidia would restrict or downgrade access to Arm's technology with regard to its rivals, undermining Arm's "neutral" position in the market.[178]

The UK's Competition and Markets Authority recently published its initial report on AI Foundation Models, which will be followed up with a subsequent workstream in early 2024. The report identifies both significant vertical integration in the AI market as well as links across the market through partnerships and strategic investments. It notes that requirements for and access to large computing power will be a key factor driving the direction of concentration in the market.[179]

Related to these concerns, the UK's Ofcom is conducting a market study into cloud services that could conclude in a recommendation to the Competition and Markets Authority to launch a formal investigation into concentration in the cloud market.[180] In its interim report, Ofcom stated that they "have reasonable grounds to suspect that there are features in the public cloud infrastructures market that may have an adverse effect on competition in the UK".[181]

Other agencies exploring antitrust concerns in cloud computing include South Korea,[182] the Netherlands,[183] Japan,[184] and France.[185]

---

[178] Federal Trade Commission, "Statement Regarding Termination of Nvidia Corp.'s Attempted Acquisition of Arm Ltd.," press release, February 14, 2022, https://ftc.gov/news-events/news/press-releases/2022/02/statement-regarding-termination-nvidia-corps-attempted-acquisition-arm-ltd.

[179] Competition & Markets Authority, "AI Foundation Models: Initial Report", September 18, 2023. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1185508/Full_report_.pdf.

[180] Ofcom, "Consultation: Cloud Services Market Study (interim report), May 17, 2023, https://ofcom.org.uk/consultations-and-statements/category-2/cloud-services-market-study.

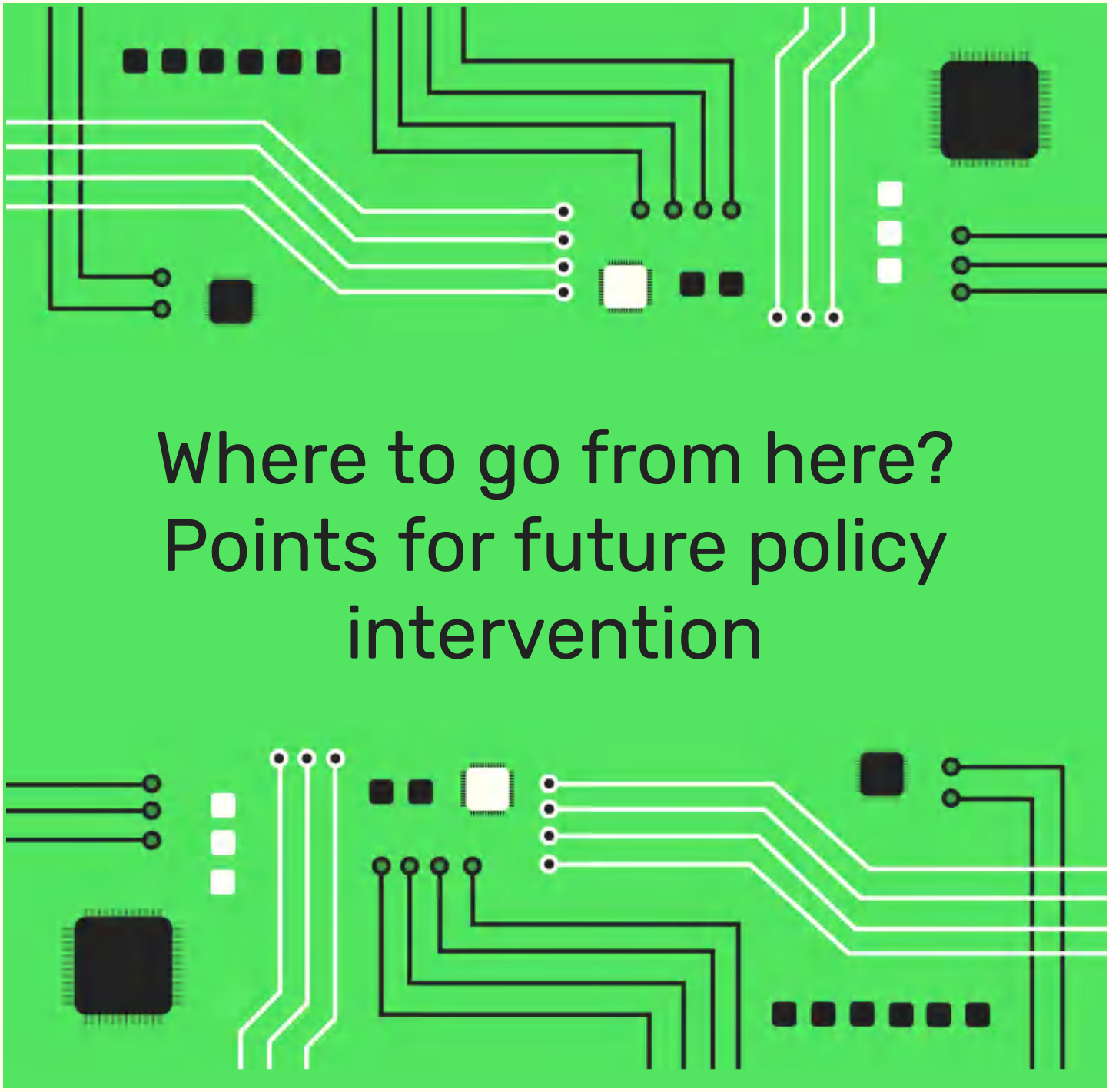
[181] Ibid.

[182] Jenny Lee, "Anticompetitive Concerns Exist in Cloud-Services Market, South Korean Antitrust Watchdog Says," MLex, December 28, 2022, https://mlexmarketinsight.com/news/insight/anticompetitive-concerns-exist-in-cloud-services-market-south-korean-antitrust-watchdog-says.

[183] Authority for Consumers & Markets, "Market Study into Cloud Services," May 9, 2022, https://acm.nl/en/publications/market-study-cloud-services.

[184] Japanese Federal Trade Commission, *Report Regarding Cloud Services*.

[185] Authorité de la concurrence, "Cloud computing: The Autorité de la concurrence Issues an Opinion on Certain Provisions of the Draft Law to Secure and Regulate the Digital Space," press release, May 12, 2023, https://autoritedelaconcurrence.fr/en/press-release/cloud-computing-autorite-de-la-concurrence-issues-opinion-certain-provisions-draft.



Where to go from here? Points for future policy intervention

Antitrust

Given the very high levels of concentration across several points in the compute stack, antitrust is a centrally relevant area for policy scrutiny. Throughout this explainer, we have seen that market participants worry about vertical integration along different parts of the

value chain:[186] AI companies worry that cloud providers will integrate downstream into AI companies' markets, and that chip providers can abuse their dominance through their bundling of software with hardware. Chip designers worry that cloud providers can integrate upstream into designing chips, and that cloud providers can abuse their dominance in the cloud market by becoming price-makers for AI chips. All Big Tech firms are now engaged in an effort to integrate a majority of the compute supply chain in-house or through exclusive agreements and strategic partnerships.

Scrutiny of these kinds of vertical integration that have started to take place in the market for AI compute is already underway. Potential regulatory approaches include the following:

1. **Separating cloud provision from chip design**

   The development of proprietary chips by cloud infrastructure providers could create vendor lock-in for the clients of those providers. This is most concerning where it deepens lock-in across an entire ecosystem by tying AI development to particular hardware and software configurations given the significant start-up costs involved in entering these markets. Separating cloud infrastructure from the design of chips themselves would ameliorate these concerns.

2. **Separating compute hardware from compute software, or mandating interoperability**

   This is best exemplified by the role of Nvidia's compiling software CUDA, which is highly specialized. Widespread familiarity with CUDA and the current lack of viable alternatives serves to reinforce Nvidia's dominant position in chip design: separating the hardware and software layers of the compute stack or mandating some level of interoperability would target this layer of vertical integration and make it easier for developers to switch between compute providers.

3. **Separating AI model development from cloud infrastructure.**

   Cloud infrastructure providers are strongly incentivized to self-preference or otherwise tilt the playing field in their favor in order to retain their dominant position in the cloud infrastructure market. Many of these providers are developing vertically integrated models of AI development, in tandem with marketplace strategies that cement their dominant hold over the ecosystem. These dual strategies have given rise to toxic competition among cloud providers, who are racing to commercially release AI systems to retain their first-mover advantage.[187] Structural separations

[186] Haydn Belfield and Shin-Shin Hua, "Compute and Antitrust: Regulatory Implications of the AI Hardware Supply Chain, from Chip Design to Cloud APIs," *Verfassungsblog* (blog), August 19, 2022, https://verfassungsblog.de/compute-and-antitrust.

[187] West and Vipra, *Computational Power and AI*.

would remove these incentive structures, while in tandem ensuring that the AI supply chain is clearly delineated in a manner that would be beneficial to other accountability and safety interventions: for example, such an intervention would make clear which entity is responsible for safety, security, design and data choices that can have significant downstream effects on AI models' behavior.

4. **Instituting nondiscrimination or common carrier obligations on compute providers operating key points of the stack**

   Compatible with structural separations, nondiscrimination obligations for cloud infrastructure or operators of other key elements of the compute stack would ensure that key infrastructures serve the interests of all customers and the broader public equally.

5. **Intervening early through merger enforcement to prevent further market concentration**

   Special attention should be paid to acquisitions of AI companies by Big Tech entities and by key gatekeepers in the compute stack such as Nvidia. The FTC has already intervened by blocking a proposed merger between Nvidia and Arm, citing potential downstream effects on the AI market, among other reasons.[188] Given the highly acquisitive behavior of firms across the tech industry and the unconventional arrangements emerging in the AI market, merger enforcement will be a key front for preventing anti-competitive behavior before harms are cemented.

6. **Investigating and tackling anticompetitive conduct**

   Swift enforcement by competition authorities against anticompetitive conduct will be key. For example, competition authorities can disallow the default bundling of some compute components, in particular compute hardware and software, and use of restrictive licensing provisions.

7. **Applying platform and marketplace-related antitrust principles to AI compute markets**

   Emerging platformization in AI compute markets can hasten vertical integration and exacerbate its harms. Important here is Amazon's Bedrock marketplace, discussed briefly above. An API service that provides access to AI models from AWS's clients, Bedrock leverages AWS's market power in the AI infrastructure market to gain market share in the AI models market, integrating across the value chain. Platform

[188] Federal Trade Commission, "Nvidia/Arm, In the Matter of."

and marketplace regulation can be applied to AI compute markets to ensure fair treatment of platform participants.

Labor Policy

- **Prohibitions on noncompete agreements would have a beneficial effect on talent bottlenecks**

  Today's high compute costs increase the cost of talent as well, because increasingly specialized talent is required to make the best of limited hardware.[189] Since limited talent can function as a barrier to entry, governments may consider interventions to reduce friction in the movement of talent across the industry, such as by prohibiting noncompete agreements, as the FTC is currently considering doing.[190]

Data Minimization

Data and compute are considered separate inputs to AI, but they are interrelated. Large amounts of good-quality data can differentiate models with the same amount of compute, and can even make some smaller models perform better than larger models with lower-quality data.[191] While scaling laws present a limit to the amount of data that can be efficiently used with a given amount of compute, exclusive data access will become increasingly important as freely available internet data runs out.[192] This is both because the freely available data will already have been used, and because newly produced data on the internet is starting to be protected more by platforms. Reddit and Twitter have already implemented some protections against free use of data from their platforms.[193] In addition, the relative value of internet data declines as the internet begins to feature more AI-generated content.[194]

---

[189] Tamay Besiroglu et al., "The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?," forthcoming.

[190] Federal Trade Commission, "FTC Proposes Rule to Ban Noncompete Clauses, Which Hurt Workers and Harm Competition," press release, January 5, 2023, https://www.ftc.gov/news-events/news/press-releases/2023/01/ftc-proposes-rule-ban-noncompete-clauses-which-hurt-workers-harm-competition.

[191] Rohan Anil et al., "PaLM 2 Technical Report," arXiv, May 17, 2023, https://doi.org/10.48550/arXiv.2305.10403.

[192] "Trends," Epoch, April 11, 2023, https://epochai.org/trends.

[193] Rohan Goswami, "Reddit Will Charge Hefty Fees to the Many Third-Party Apps That Access Its Data," CNBC, June 1, 2023, https://cnbc.com/2023/06/01/reddit-eyeing-ipo-charge-millions-in-fees-for-third-party-api-access.html.

[194] Ilia Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget," arXiv, May 31, 2023, https://doi.org/10.48550/arXiv.2305.17493.

AI companies that cannot outspend competitors on compute will naturally try to achieve relative quality improvements through data. This incentive increases even for the largest spenders on compute as models become larger. In this context, it is important to note that cloud providers in particular and Big Tech in general have exclusive access to vast repositories of personal and nonpersonal data. OpenAI has reportedly already used YouTube data to train its models, which leaves the door open for Google to use data not only from YouTube, but also from Gmail, Google Drive, and all its other services.[195] Similarly, Microsoft can potentially use data from its enterprise services, and AWS from its cloud services. However, data is an increasingly opaque element of the market and many companies are not willing to disclose what data they are using to train their models.

A separations regime as described above can prevent some of this wanton data use, but laws targeted toward data protection and sharing can go further. At the very least, clarifications on the legality of using data from other services to train AI models can deter such use. More watertight enforcement of the law will require monitoring mechanisms. For instance, recent work shows that the data used to train a model can potentially be verified.[196] Governments can also act to further protect domain-specific data, including healthcare and education data where a concentrated AI market can be especially damaging.

Governments can thus prevent further concentration of the AI compute market, not to mention the AI models market, by updating and enforcing data protection rules, including:

- **Embracing data-minimization mandates, including prohibiting the collection or processing of all sensitive data beyond what is strictly necessary**

  Data policy *is* AI policy, and curbs on unbridled commercial surveillance practices will have important effects on AI. The FTC in particular should consider secondary use of commercial surveillance data as part of its Advance Notice of Proposed Rulemaking (ANPR) on surveillance and data security, as this relates to data practices that potentially harm consumer protection and competition.[197]

- **Prohibiting the secondary use of data collected from consumers for the purpose of training AI models, as a violation of consumer control over personal data**

[195] Jon Victor, "Why YouTube Could Give Google an Edge in AI," *The Information*, June 14, 2023, https://theinformation.com/articles/why-youtube-could-give-google-an-edge-in-ai.

[196] Dami Choi, Yonadav Shavit, and David Duvenaud, "Tools for Verifying Neural Models' Training Data," arXiv, July 2, 2023, https://doi.org/10.48550/arXiv.2307.00682.

[197] Federal Trade Commission, "Commercial Surveillance and Data Security Rulemaking," August 11, 2022, https://http://ftc.gov/legal-library/browse/federal-register-notices/commercial-surveillance-data-security-rulemaking.

The FTC has already outlined this principle in its recent Amazon Alexa case.[198] Since data is an important differentiator for AI model quality, cloud infrastructure providers may be engaging in anticompetitive behavior if they utilize user data to train AI models. They may also be violating well-established data protection principles like purpose limitation.

[198] Federal Trade Commission, "FTC and DOJ Charge Amazon with Violating Children's Privacy Law by Keeping Kids' Alexa Voice Recordings Forever and Undermining Parents' Deletion Requests," press release,, May 31, 2023, https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-doj-charge-amazon-violating-childrens-privacy-law-keeping-kids-alexa-voice-recordings-forever.