Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe

Vili Lehdonvirta[1,2], Bóxī Wú[1], Zoe Hawkins[3]

[1]Oxford Internet Institute, University of Oxford, Oxford, UK
[2] Department of Computer Science, Aalto University, Espoo, Finland
[3] Tech Policy Design Centre, Australian National University, Canberra, Australia
vili.lehdonvirta@oii.ox.ac.uk, boxi.wu@jesus.ox.ac.uk, zoe.hawkins@anu.edu.au

Abstract

Governments have begun to view AI compute infrastructures, including advanced AI chips, as a geostrategic resource. This is partly because "compute governance" is believed to be emerging as an important tool for governing AI systems. In this governance model, states that host AI compute capacity within their territorial jurisdictions are likely to be better placed to impose their rules on AI systems than states that do not. In this study, we provide the first attempt at mapping the global geography of public cloud GPU compute, one particularly important category of AI compute infrastructure. Using a census of hyperscale cloud providers' cloud regions, we observe that the world is divided into "Compute North" countries that host AI compute relevant for AI *development* (ie. training), "Compute South" countries whose AI compute is more relevant for AI *deployment* (ie. running inferencing), and "Compute Desert" countries that host no public cloud AI compute at all. We generate potential explanations for the results using expert interviews, discuss the implications to AI governance and technology geopolitics, and consider possible future trajectories.

1 Introduction

Compute, data, and algorithms are three core inputs in the development and deployment of artificial intelligence (AI) systems (Sevilla 2022). Of these, compute has emerged as a particularly important resource over the last few years, as the computing power used to train frontier AI models has doubled approximately every six months (Heim et al. 2024). Where in the world is all the compute used in AI development and deployment physically located? We argue that this question is becoming increasingly important from the perspectives of AI governance and geopolitical analysis.

Sastry and colleagues (2024) argue that "compute governance" is emerging as an important tool for governing the development and deployment of AI systems. This is because compared to data and algorithms, compute is an observable and measurable aspect of AI training which can be restricted through the physical nature of compute hardware (Sastry et al. 2024). The compute used to train and deploy a large-scale AI model is typically produced by hundreds or even thousands of graphics processing units (GPUs) and other AI accelerator chips, which are housed inside large data centers that consume megawatts of electric power (Pilz and Heim 2023). Regulators could for instance implement legal compliance checks at the point at which algorithms and data arrive at a data centre for the purpose of enabling shared security standards, government record keeping, and the verification of AI systems and developers, or to enforce restrictions and limitations on non-compliant systems (Heim et al. 2024).

Compute governance could potentially empower states that don't have other means to enforce AI regulations (Radu 2021; Png 2022), such as jurisdiction over the developers (as the United States), or jurisdiction over important market areas for the developers (as the European Union). However, not all countries are equally able to implement governance via compute. States that can exercise territorial jurisdiction over physical compute infrastructures are better placed to impose their rules on them than states that cannot (Ferrari 2024). This is because even if a state's laws are intended to cover the contents of computational infrastructures situated abroad, the enforcement of those laws can be difficult and become contested by other states (Abraha 2019; 2021).

For similar reasons, the geography of AI compute has also become a subject of geopolitical contestation. Many governments see AI as a potential source of national economic, military, and/or cultural advantage (Wang and Chen 2018; Miller 2022). The United States, China, United Kingdom, United Arab Emirates, and many other governments have adopted industrial policies aimed at securing sufficient supplies of AI compute for local industries and researchers (Vipra and West 2023). The United States has also adopted a policy of restricting China's access to compute in an effort to contain its AI progress (Allen 2022). Following the *Creating Helpful Incentives to Produce Semiconductors*

(CHIPS) Act in 2022, the United States (U.S.) government restricts exports of advanced GPUs to China. The government is also considering imposing restrictions on the remote provision of cloud computing services to Chinese AI developers (Edgerton 2023). Chinese developers would then only have access to domestic compute and possibly to compute hosted in sympathetic nations. Insofar as AI development is important for national economic, military, and/or cultural competitiveness, the physical location of compute infrastructures thus matters in geopolitical analysis.

The purpose of our study is to provide the first attempt at mapping the global geography of public cloud GPU compute, one particularly important category of AI compute infrastructure. We achieve this by carrying out a census of the cloud regions of six leading hyperscale public cloud computing providers: Amazon Web Services (AWS), Microsoft Azure, Google Cloud, Alibaba Cloud, Huawei Cloud, and Tencent Cloud (Lehdonvirta, Wu and Hawkins 2023). We also examine potential explanations for the observed geographies based on expert interviews and discuss the implications for compute governance and geopolitics.

2 Background and Related Work

The global AI compute supply chain broadly speaking consists of (1) companies that design and market GPUs and other AI-relevant chips, (2) companies that fabricate and package the chips, (3) companies that deploy the chips to provide compute, and (4) companies that consume the compute to develop and/or deploy AI systems (OECD 2023). Previous work on the geography of compute has focused especially on the design and fabrication steps (Miller 2022). The market leader in GPU design and marketing is U.S.-based Nvidia corporation, fabrication is dominated by Taiwanese TSMC, and Dutch ASML is currently the sole producer of photolitography machines that are essential in the fabrication of the most advanced chips (Miller 2022). These parts of the compute supply chain are thus very concentrated in terms of both geography and ownership. Much of the previous literature discusses the economic and political implications of this concentration (Allen 2022; Yeung 2022), as well as attempts by governments to shape it to their advantage (Drezner, Farrell and Newman 2021; Farrell and Newman 2019; 2023), such as the U.S. CHIPS Act and export controls aimed at restricting the export of GPUs to China (Heim et al. 2024; Sastry et al. 2024) .

In this study, we are concerned with the geography of the third step: Where in the world are chips deployed to provide AI compute for AI development and deployment, that is, to train AI models and run inferencing on existing models? There are broadly speaking three types of compute providers doing this on a large scale: scientific supercomputing facilities, private compute clusters, and so-called public cloud providers.

Scientific supercomputing facilities have existed from the early 1960s. They are typically government-funded and intended mainly for academic and military use. A study by OECD (2023) provides a simple geographic analysis of scientific supercomputing facilities. The highest concentration of supercomputers listed in the TOP500 database is found in China (32%), followed by the United States (25%) and the European Union (21%). However, most scientific supercomputers were not designed with AI model training in mind (OECD 2023). The current generative AI development boom has been powered mostly by private compute clusters and by public cloud compute. Previous research has not attempted to analyse their geography in any detail.

Private compute clusters are owned by for-profit companies, such as Meta, HP, and many smaller firms. They consist of interconnected GPU-equipped computers deployed in data centres. A private cluster can be used to power the company's own AI development or rented out to another company. Public cloud providers are likewise for-profit companies. They are called "public" not because of any government affiliation but because their services are available on demand and shared by many customers (i.e. the public in public house rather than in public sector) (Herr 2020). The market leaders in public cloud computing, including public cloud AI compute are AWS, Microsoft Azure, and Google Cloud; Chinese public cloud providers Alibaba, Huawei, and Tencent also provide AI compute at scale (Lehdonvirta, Wu and Hawkins 2023). These large providers are often referred to as "hyperscalers" (Vipra and West 2023).

In this study we choose to focus on the geography of public cloud AI compute. Private compute clusters have been used to train some landmark models such as Meta's Llama and Llama 2 (Sastry et al. 2024). But much of the training and development of frontier AI models is concentrated within public cloud hyperscalers Google, Microsoft, and Amazon, and their corresponding "compute partnerships" with leading AI companies such as Anthropic, Cohere, Google DeepMind, Hugging Face, OpenAI, and Stability AI (Sastry et al. 2024). Public cloud is also important because it is accessible to a great number and variety of developers, including academic researchers. Our main research question therefore is, where in the world is public cloud AI compute located? We will also examine potential explanations for the observed geographies, discuss their implications for compute governance and geopolitics, and finally briefly discuss private clusters and government-owned national AI compute.

It is worth noting that any effort to map or measure compute is always imprecise in some way. Although the term

"compute" is often used in a way that suggests it is a fungible commodity, measurable in units such as FLOPs (floating point operations), in practice computation is heterogeneous: chips and system architectures that excel at one type of task may perform significantly worse in another. In this study, we use the term "AI compute" to refer to computational capabilities that are particularly relevant for tasks related to today's AI systems. A further distinction will be made between compute relevant for AI model development (training) and compute more relevant in AI system deployment (running inferencing).

3 Methodology

Our methodology is two-fold. We first conduct a census of public cloud providers' cloud regions and analysed the geographic distribution of regions where customers can access GPUs. We also interviewed 10 experts from policy, academia and the public cloud providers themselves to develop potential explanations.

3.1 Census of Public Cloud Regions

Our census comprised the six hyperscale public cloud providers mentioned earlier: AWS, Microsoft, Google, Alibaba, Huawei, and Tencent. Although there are also smaller providers, these six represent the great majority of the global public cloud market as well as each regional market (Lehdonvirta, Wu, and Hawkins 2023). We used providers' websites and customer interfaces to collect a list of each provider's public cloud regions as of October 2023. A "cloud region" is essentially a cluster of interconnected data centres and supporting infrastructure located in a specific geographic area, named after a nearby city (Google, n.d.). To map the region names to approximate physical locations we relied on the World Cities Database (World Cities Database, 2024). A developer wishing to use public cloud AI compute must choose the region they would like it to be physically hosted in. Different regions may support different types of GPUs or none at all. We collated this information into a database, using the variables shown in Table 1.

Variables	Levels
Provider	6 (AWS…Tencent)
Country	39 (Argentina…United States)
H100 available	2 (Yes/No)
A100 available	2 (Yes/No)
V100 available	2 (Yes/No)
H800 available	2 (Yes/No)
A800 available	2 (Yes/No)
N of cloud regions	187

Table 1. Public cloud region census 2023 variables (unit of analysis: cloud region)

When the census was taken, the most powerful GPU for the purposes of training common AI models was the Nvidia H100, launched in 2023 (Pilz and Heim 2023). The previous flagship model A100 was launched in 2020, and the V100 before that in 2017. In 2023 Nvidia introduced the H800 and A800 to circumvent U.S. export restrictions to China, but the restrictions were quickly expanded to cover them (Reinsch, Schleich, and Denamiel 2023). We focused our data collection on these five most AI-relevant GPU types (Table 2). Custom AI accelerator chips such as the Google Tensor Processing Unit (TPU) were excluded from the census for logistical reasons, which is a limitation in our study.

	Mean	SD	Min	Med	Max
Total cloud regions	4.7	6.7	1	3	36
GPU-enabled regions	2.5	5.3	0	1	27
H100-enabled regions	0.2	1.3	0	0	8
A100-enabled regions	1.2	3.1	0	0	18
V100-enabled regions	2	3.9	0	1	19
H800-enabled regions	0	0	0	0	0
A800-enabled regions	0	0	0	0	0
N of countries	39				

Table 2. Public cloud AI compute infrastructure by country (unit of analysis: country)

From the census database we then constructed a country-level data set to facilitate geographic analysis. For each country we calculated the total number of public cloud regions situated in its territory. We also calculated the subset of regions that supported at least one type of GPU ("GPU-enabled region"), as well as the subsets of regions that supported specific GPU types. In practice we did not observe any H800s or A800s being made available by public cloud providers in China or elsewhere. Although Hong Kong is a special administrative region of China, it is treated as a separate entity in this data set, because it has a partly separate legal system and because public cloud providers distinguish between mainland Chinese and Hong Kong regions.

A notable weakness in this approach to mapping public cloud AI compute is that it does not consider how many GPUs of each type each of the cloud regions makes available to customers. Nor does it consider other system architecture parameters, such as memory or maximum cluster size. The approach simply counts regions and their capabilities in terms of GPU types available. We partially address these limitations with the expert interviews.

3.2 Expert Interviews

To complement our quantitative cloud census, we carried out a series of qualitative and semi-structured expert interviews (Babbie 2016; Mikecz, 2012). We interviewed a total of ten informants representing two policy, three hyperscale

public cloud provider, and five research informants with expertise in AI compute (Table 3). The informants were recruited by snowball sampling through our own professional networks.

Informant	Interview date	Stakeholder
A	8 Feb 2024	Policy
B	2 Apr 2024	Policy
C	5 Feb 2024	Research
D	18 Oct 2023	Research
E	19 Jan 2024	Research
F	5 Dec 2023	Research
G	19 Jan 2024	Research
H	11 Jan 2024	Hyperscaler
I	15 Feb 2024	Hyperscaler
J	15 Dec 2023	Hyperscaler
N of informants	10	

Table 3: Overview of informants

The main objectives of the interviews were to help improve and validate our census approach, to generate complementary or alternative information on the geographic distribution of public cloud AI compute, and to help generate explanations for the geographic patterns observed. The interviews took place between December 2023 and March 2024. They were carried out via both video conferencing and in person meetings, and followed a semi-structured approach with questions designed to address a set of topics in-depth that reflected our objectives (Babbie, 2016). Interview recordings were transcribed and the research team performed a simple thematic analysis of the transcripts to surface answers on the issues of interest.

In the following sections we report quantitative findings from the cloud census interspersed with interpretations and complementary answers surfaced from the interviews.

4. Where is AI Compute Located?

Figure 1 shows the approximate locations of the public cloud regions discovered in our census. Table 4 indicates how many of the regions are located in each country and how many of those regions offer GPU instances ("GPU-enabled regions"). Arguably the most important feature of the data from the point of view of compute governance is that the vast majority of countries in the world have no public cloud regions at all. Of the 39 countries that do have one or more cloud regions, 30 have cloud regions that feature GPUs.

Another striking feature of the data is that even within those countries that host some GPU-enabled cloud regions, the geographic distribution of the regions is very polarized: China and the United States together host almost as many regions (49 regions) as the rest of the world put together (52 regions). Of the two, China's total number of GPU-enabled regions is slightly higher (27) than the US (22).

Further insight may be gained by examining what types of GPU instances each country is hosting. The most obvious pattern is that the U.S. is hosting the newest and most powerful GPUs in the world both in terms of the ratio of different types of instances available and in absolute numbers. The U.S. is the only country to have more regions offering the 2020 Nvidia A100 GPU than the 2017 V100 GPU. The U.S. also has multiple regions offering the 2023 Nvidia H100 GPU. China's cloud regions are mostly based on the V100, with a smaller number of regions offering A100s. No regions in China offer the H100. Across the rest of the world only 15 countries offer A100s and only one has H100s, the rest being purely V100-based.

As noted, this analysis does not consider custom accelerator chips such as TPUs, nor potential differences in the quantities of GPUs of different types available in different regions. Our interview informants noted that the quantity of GPUs of a given type available in a region is likely to vary significantly between regions and providers. "Hyperscalers have succeeded in giving the impression that they are almost omnipotent when it comes to compute or storage, that they can handle anything you bring. But that's not quite the reality", noted one informant. In some cases the quantities of GPUs available in a region can be very limited, with the consequence that only a limited number of customers can run GPU instances in that region, and/or only models of a limited size can be trained in a reasonable time span in that region.

AWS and Microsoft are currently thought to have the largest cloud GPU clusters available, but "regions are definitely not identical in this respect." That said, GPU quantities and especially how they are distributed across the providers' regions are treated as highly confidential information by hyperscale cloud providers. None of our informants were willing or able to give us any concrete numbers nor identify how this information could be publicly accessed. But it was generally agreed that U.S.-based regions were likely to have larger quantities of GPUs of any given type than regions elsewhere in the world featuring the same GPU type. Chinese regions might also feature V100 chips in larger numbers to make up for their comparatively lower performance. Our interviews thus suggest that even if it was possible to include the quantity of GPUs per region in this analysis, it would probably not challenge the major patterns noted above, but more likely amplify them.

Country	Total regions	Total GPU-enabled regions	H100-enabled regions	A100-enabled regions	V100-enabled regions
Argentina	1	1	0	0	1
Australia	9	3	0	1	3
Bahrain	1	0	0	0	0
Belgium*	1	0	0	0	0
Brazil	5	1	0	0	1
Canada	6	2	0	1	2
Chile	2	1	0	0	1
China	36	27	0	9	19
Finland*	1	0	0	0	0
France*	4	1	0	1	1
Germany*	5	2	0	1	2
Hong Kong	6	3	0	1	3
India	9	3	0	1	2
Indonesia	5	2	0	0	2
Ireland*	3	2	0	2	2
Israel	1	1	0	1	0
Italy*	5	1	0	1	0
Japan	9	4	0	3	3
Korea	5	4	0	1	3
Malaysia	1	1	0	0	1
Mexico	2	2	0	0	2
Netherlands*	4	2	1	2	2
Norway	1	0	0	0	0
Peru	1	1	0	0	1
Philippines	1	0	0	0	0
Poland*	2	0	0	0	0
Qatar	2	0	0	0	0
Saudi Arabia	2	1	0	0	1
Singapore	6	5	0	2	4
South Africa	3	1	0	0	1
Spain*	2	0	0	0	0
Sweden*	2	1	0	1	0
Switzerland	3	1	0	0	1
Taiwan	1	1	0	0	1
Thailand	3	2	0	0	2
Turkey	1	1	0	0	1
UAE	3	0	0	0	0
UK	6	2	0	2	2
US	27	22	8	18	17
**EU27 total*	*34*	*11*	*1*	*8*	*9*
Global Total	*187*	*101*	*9*	*48*	*81*

Table 4. Public cloud regions by country

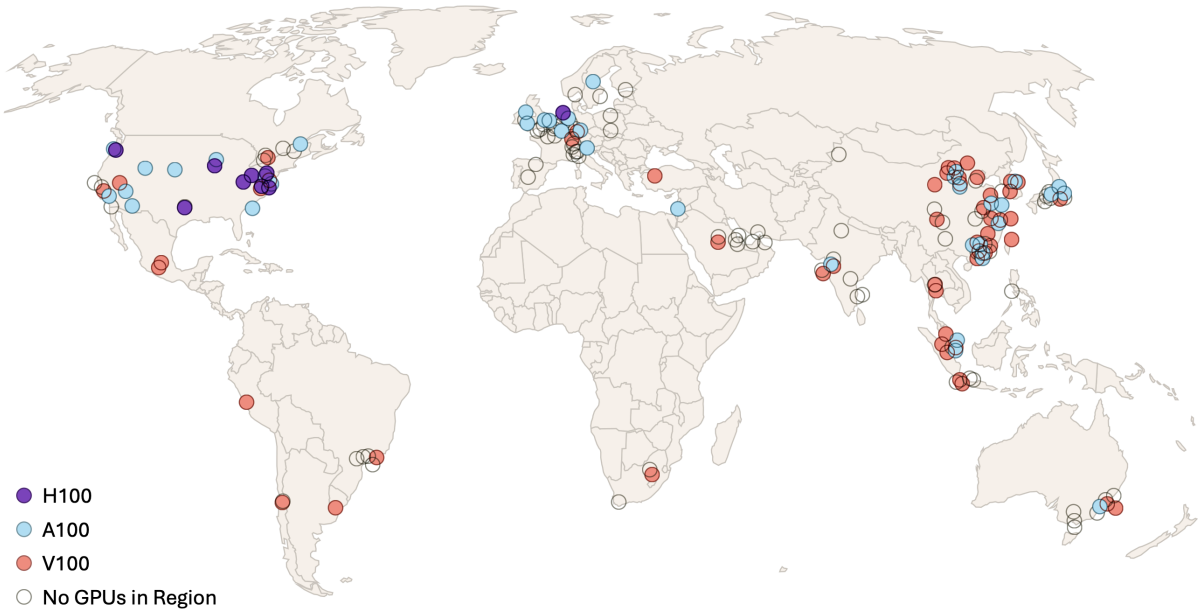

Figure 1. Approximate locations of public cloud regions and the most advanced GPU type available in each region

5 Why is Advanced AI Compute Concentrated in the US?

What explains the apparent U.S. lead in advanced public cloud AI compute over China and other countries? One obvious explanation is U.S. government export controls that forbid exports of A100s and H100s to China (Reinsch, Schleich, and Denamiel 2023). Chinese cloud providers were able to import some A100s before the export controls took effect in 2023, but H100s have been export controlled from since the product was launched. The H800s and A800s were likewise export controlled soon after introduction. The considerably less powerful V100, which is the most common Nvidia GPU instance type in China, is not subject to export controls. One informant argued that if more powerful GPUs were finding their way into China via grey imports, then they were probably being deployed in private clusters where they could fly under the radar, as public cloud providers openly circumventing the controls could jeopardize their ability to do business with Nvidia on non-controlled chips.

Yet export controls cannot explain why other countries besides China also host predominantly older GPUs. Several explanations are possible. One is simple friction in innovation diffusion, referring to the process through which GPUs spread throughout markets (Rogers 2010). It may be that newer GPUs are first installed in the U.S. since that is where Nvidia is based and therefore where it has the strongest distribution networks. Over time advanced GPUs should diffuse to markets that are relatively farther away. "I would assume that almost all [GPUs] come to North American regions first, but by now there should be significant clusters in Europe as well," one informant speculated.

Another potential explanation for the U.S. cloud compute lead stems from geographic differences in initial demand profiles, which combines with scale economies to create a "path dependency" that sustains the concentration of AI compute to certain geographies (Radu and Amon 2021). One informant explained:

> "Very few [cloud compute buyers] are actually doing groundbreaking AI development… So there's no point spreading the capacity everywhere… you want a few super units, you want to have a critical mass of compute in some location, which is not worth replicating everywhere."

The first significant concentration of firms and researchers training large AI models emerged in the U.S., so cloud providers concentrated their most powerful training compute capacity there. But even as the demand for compute grew in other locations around the world, this did not necessarily translate into corresponding growth in local compute infrastructure, because developers could usually send their training workloads to U.S. cloud regions without incurring significant performance penalties. The initial U.S. compute lead was thus sustained.

Informants argued that the situation was somewhat different for AI compute capacity intended for deploying rather than training AI. In many AI use cases, such as with voice assistants, the user experience can suffer from latency if the distance between the user and the server is too great. Data transfer costs may also become a business issue. Such appli-

cations are thus best deployed on compute infrastructure situated physically closer to users. This would explain why less advanced V100 chips—that may be too slow for training purposes but still adequate for inferencing—are distributed more evenly around the globe than more advanced chips.

Some exceptions are evident to the general pattern of the U.S. having the most advanced GPU assortment. Japan, the United Kingdom, and France each host the same number of A100-enabled regions as they host V100-enabled regions. Each of these countries has significant local AI development activity. There may be regulatory or political obstacles to local developers sending data to the U.S. for training (Komaitis 2017; Herr 2020). According to one informant:

> "By now there are public sector actors or significant European actors that have the need to train GPT-4 level models with data that cannot be taken outside Europe… I would be surprised if hyperscalers weren't responding to that demand."

In this context, informants referred to policy discourse on "digital sovereignty", "data sovereignty", and "compute sovereignty" as potentially creating demand for locally situated training compute (Tang 2022; Pohle and Thiel 2020; Gu 2023; Roberts et al. 2023). The Netherlands and Ireland also have small but relatively advanced GPU assortments. This is perhaps related to these countries' strategic positions as infrastructure hubs for some of the hyperscalers (Herr 2020; Rone 2023). Netherlands is particularly notable as the only country besides the U.S. to host a cloud region featuring the powerful H100 GPU.

6 Discussion

6.1 A Global Compute Divide

Governing AI through compute is a powerful idea, because compute is made up of large, observable material infrastructures (Sastry et al. 2024; Heim et al. 2024). The infrastructures must be physically situated somewhere, and are therefore susceptible to territorial jurisdiction, the most enforceable form of jurisdiction that all states—small and large—possess (Mikler 2018). However, our research shows that compute infrastructures are not situated evenly across the globe, and their geographic distribution strongly conditions different states' possibilities of turning compute into a point of intervention into AI.

Our findings reproduce the familiar idea of two AI superpowers engaged in a compute "arms race" (Wang and Chen 2018), in which the U.S. holds an edge in terms of chip quality and China attempts to compensate with quantity (Miller 2022). The U.S. export restrictions on advanced GPUs appear to have worked, as no public cloud providers offer the 2023 H100 chip in China, nor the H800 or A800 created to circumvent the restrictions. Similarly, Russia and Iran, two countries subject to Western sanctions, do not host any public cloud AI compute belonging to the providers in our sample.

However, beyond ideas of geopolitical great power rivalry, our findings also suggest additional conceptual categories relevant to discussing compute-based AI governance. There are 15 other countries besides the U.S and China that also host at least some quantity of the GPUs most relevant to AI development, namely A100s and H100s. All of these first-tier countries save for India are located in the so-called Global North. To draw an analogy, we refer to them as the "Compute North" (Figure 2). These Compute North countries are positioned to use their territorial jurisdiction to intervene in AI development at the point at which models are sent to their local public cloud regions for training. For instance, they could require algorithms and data sets to be audited and certified for compliance with their local rules before training is permitted to commence, shaping what kinds of AI systems can enter the global market.

A second tier of 13 countries hosts compute of a type more suited for AI system deployment than for development. All of these countries are situated in the Global South, save for Switzerland; we refer to them as the "Compute South". For instance, Latin America hosts a total five GPU-enabled cloud regions, but none of them featured more powerful GPUs than the 2017 V100. These countries are positioned to use their territorial jurisdiction over compute to gatekeep which AI systems can be deployed locally, but less so for shaping AI system development.

Besides the "Compute North" and the "Compute South", there is also a "Compute Desert", by which we refer to all of the remaining countries in the world. These countries host no public cloud AI compute at all, whether for training or for deployment. For them, shifting to cloud-based AI-powered services means relying on services both developed and deployed on infrastructures located in foreign jurisdictions. The Compute Desert contains a number of rich countries, but it also contains all of the world's lower middle-income and lower-income countries, following the International Monetary Fund's (IMF) classification (IMF 2024). The implications of belonging to the Compute Desert are likely to differ between rich and poor countries. The rich countries in the Desert may be able to use their other advantages—such as diplomatic influence over Compute North countries, and wealth sufficient to build government-owned compute capacity—to offset their lack of locally hosted public cloud AI compute (Png 2022; Ferrari 2024). In contrast, the poorer countries in the Compute Desert have few prospects for making use of compute governance as a means to influence AI.

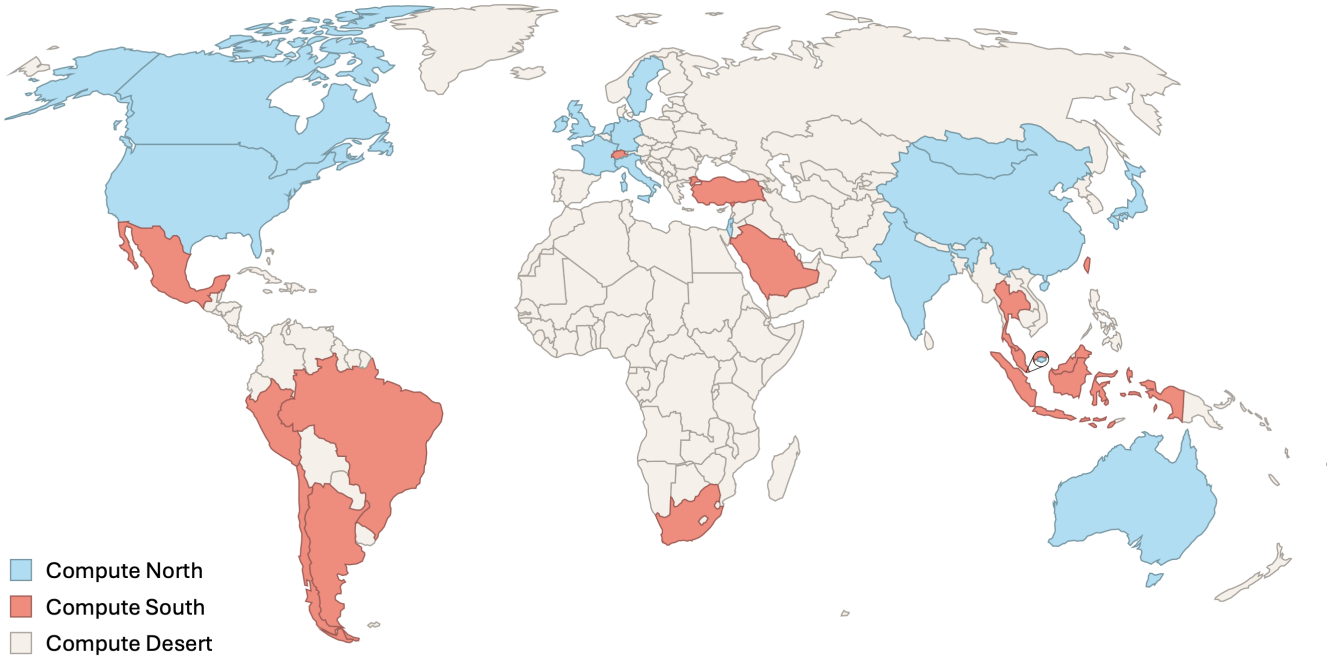

Figure 2: Global compute divide between the Compute North, Compute South, and the Compute Desert, based on public cloud GPU compute

Similar to how researchers have observed a "compute divide" between academia and industry (Besiroglu et al. 2024), we thus observe a global compute divide, in which the geography of public cloud AI compute seems to be reproducing familiar patterns of global inequality. From mid-1990s, discourses on digitalization posited that success in the new global "knowledge economy" would be based on immaterial assets such as knowledge and creativity, in contrast to material assets and resources required in the previous era of industrial economies (Negroponte 1995; Castells 1998). This would allow developing countries to forgo expensive infrastructure investments and "leapfrog" directly into a knowledge-based economic model. However, today's AI discourse has once again returned to emphasizing material infrastructures such as chip fabs, data centres, and electricity networks as crucial for national competitiveness (Miller 2022; Vipra and West, 2023). If compute becomes a critical governance node, then such material infrastructures may turn out to be essential for retaining independent regulatory agency as well (Lehdonvirta 2023). A nation's computational power then equals, to an extent, its political power.

Can the situation be expected to change? If the observed concentration of high-end AI compute to the U.S. and to the Compute North is explained simply by frictions in innovation diffusion, then it is plausible that over time the globe could be increasingly saturated with compute, evening out the disparities. Nvidia's competitors such as AMD and Intel are catching up in chip performance. Chinese Huawei is also developing AI processing chips, and it is backed by enormous domestic demand following the US export controls, as well as Chinese government support (He 2024; The Economist, 2024).

But if the geographic patterns we observed are explained more by path dependence resulting from first-mover advantages and scale economies, then it is possible that geographic concentration, regional specialization, and international divisions of labour will be enduring features of compute production, as they have been in many other industries.

6.2 Public Cloud vs. Private and Government Compute Globally

In this study we focused on public cloud compute, an important but by no means the only source of compute. Within public cloud compute, our data collection was targeted at Nvidia's GPUs and six leading hyperscale cloud providers.

Could the relative standing of different types of large-scale compute providers be expected to change and challenge the observed geography of compute? An expensive capital good such as a GPU cluster needs a high utilization rate to achieve reasonable returns to investment, which explains why large clusters have been built mainly as shared infrastructures, whether government owned as in the case of scientific supercomputing, or more recently privately owned

as in the case of public cloud (Herr 2020). Government-owned compute now appears to be experiencing a small comeback in the form of "national AI compute" initiatives announced around the world (OECD, 2023). For example, the U.S. National AI Resource (NAIR) Task Force aims to create public compute infrastructure to "democratize AI research" (NAIR, 2023; Vipra and West 2023). However, the scale of government investment in many cases does not appear sufficient to seriously challenge the dominance of hyperscale cloud providers. Many recent government efforts are also undertaken in collaboration with the hyperscalers, so that in practice projects rely on privately owned infrastructure (Vipra and West 2023; Tardieu 2022).

A counterpoint is provided by the new LUMI supercomputer of the European High-Performance Computing Joint Undertaking, a collaboration of EU member country governments. Located in Kajaani, Finland, LUMI comprises a cluster of 11,912 GPUs designed by Nvidia's competitor AMD (Brans 2024). At that scale, it may provide a serious alternative to privately owned "public" cloud computing infrastructures as AI development infrastructure. Given that it is located in the EU, it does not challenge the North-South compute divide illustrated in Figure 2. It may, however, contribute to complicating the bipolar image of the U.S. and China as the only AI superpowers.

New private compute clusters also appear to be growing. Google's TPUs likely represent a significant fraction of all AI compute. AWS and Microsoft both have plans to produce their own chips. Meta announced a very significant investment into building up their private compute capacity: CEO Mark Zuckerberg claimed to be investing in 340,000 Nvidia H100s and A100s (Heath 2024). In 2023, Microsoft claimed to have spent hundreds of millions of dollars on a cluster to power OpenAI's ChatGPT chatbot (Roth 2023). Massive tech companies may be able to achieve high utilization rates for large clusters simply with their internal and partner demand. But clusters initially deployed as private can later be turned into shared cloud infrastructure once their novelty has worn off and internal demand subsided. This blurs the difference between private and public (as in public house) cloud compute capacity.

6.3 Future Directions for Research

The most obvious extensions of this project entail expanding the scope of the mapping effort to broader varieties of compute providers and to more types of AI accelerators and architectures, potentially with new data sources. But the existing findings also point to potentially important research questions in the landscape of compute governance. Researchers could seek to investigate what factors explain a country's position across the Compute North, Compute South, and Compute Desert categories. Although the categories were broadly correlated with national income levels, there were notable exceptions. Can some differences be attributed, for instance, to government policies?

Research and policy understanding on privately owned compute infrastructures is currently hindered by the opaqueness and lack of transparency of their owners, which stems in part from business confidentiality reasons and the increasingly competitive dynamics of the AI industry (Vipra and West 2023). But insofar as AI compute is also becoming a governance lever, the secrecy around compute infrastructures is out of step with the public interest. Norway has recently made it obligatory for data centre operators to register with the government and provide basic details on their capabilities and workloads. As Trager et al. (2023), Heim et al. (2024) and Sastry et al. (2024) have argued, we can expect more governments to begin to intervene on compute with transparency, disclosure, and registration requirements. As a key input into AI, the management or monitoring of compute resources as well as the implementation of multilateral controls will likely be a core concern of broader international AI governance efforts (Ho et al. 2023; Trager et al. 2023). This may open up new opportunities for researchers to understand the global geography of compute in greater detail and shape the direction of AI research in terms of academic compute access and outside scrutiny of private sector activity (Besiroglu et al. 2024).

One important issue that we did not consider in this study is that the nationality of hyperscale cloud computing providers also matters for governance and geopolitics. Providers must comply with the laws of the jurisdictions in which their global infrastructures are physically located, but they also have to comply with the laws of their home countries. Sometimes the latter laws have extraterritorial intent, such as the U.S. Cloud Act of 2018, which allows U.S. law enforcement agencies to issue warrants for data held by U.S. companies regardless of where in the world the data centre is physically located (Abraha 2019; 2021). China's National Intelligence Law may likewise have extraterritorial effects. The home states of the major cloud providers could in principle attempt to use their private companies as geopolitical tools to project power globally (Tang 2022; Gjesvik 2023; Gu 2023). Future research on compute governance and geopolitics should thus consider not only the physical location but also the nationality of AI compute infrastructures.

Acknowledgements

This study was supported by a grant from the Dieter Schwarz Stiftung. The authors gratefully acknowledge helpful comments from Lennart Heim, Lewis Ho, Jukka-Pekka Ahonen and three anonymous peer reviewers. The visual design of Figure 1 and Figure 2 is by Juuso Koponen.

References

Abraha, H.H. 2019. How compatible is the US 'CLOUD Act' with cloud computing? A brief analysis. *International Data Privacy Law* 9(2): 207-215. https://doi.org/10.1093/idpl/ipz009.

Abraha, H. H. 2021. Law enforcement access to electronic evidence across borders: Mapping policy approaches and emerging reform initiatives. *International Journal of Law and Information Technology* 29(2): 118–153. https://doi.org/10.1093/ijlit/eaab001.

Allen, G. C. 2022. Choking off China's Access to the Future of AI. *Centre for Strategic and International Studies*. https://www.csis.org/analysis/choking-chinas-access-future-ai. Accessed: 2024-05-01.

Babbie, E. R. 2016. *The Practice of Social Research*. Fourteenth edition. Boston, MA: Cengage Learning.

Besiroglu, T.; Bergerson, S. A.; Michael, A.; Heim, L.; Luo, X.; and Thompson, N. 2024. The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? arXiv:2401.02452.

Brans, P. 2024. Could Display Technologies Eliminate Bottlenecks in HPC?. https://www.eetimes.eu/could-display-technologies-eliminate-bottlenecks-in-hpc/. Accessed: 2024-05-01. Accessed: 2024-05-01.

Castells, M. (1998). *End of millennium, the information age: Economy, society and culture (Vol. III).* Blackwell.

Drezner, D. W.; Farrell, H.; and Newman, A. L. (Eds.). 2021. *The Uses and Abuses of Weaponized Interdependence*. Brookings Institution Press.

Edgerton, A. 2023. US Said to Consider Limits on Cloud Computing For China. Bloomberg. https://www.bloomberg.com/news/articles/2023-07-05/us-said-to-consider-limits-on-cloud-computing-for-china. Accessed: 2024-05-01.

Farrell, H. and Newman, A. 2019. Weaponized Globalization: Huawei and the Emerging Battle over 5G Networks. *Global Asia* 14 (3): 8–12.

Ferrari, F. 2024. State Roles in Platform Governance: AI's Regulatory Geographies. *Competition & Change* 28 (2): 340–58. https://doi.org/10.1177/10245294231218335.

Google. Geography and Regions | Documentation | Google Cloud. https://cloud.google.com/docs/geography-and-regions. Accessed: 2024-05-11.

Gjesvik, L. 2023. Private Infrastructure in Weaponized Interdependence. *Review of International Political Economy* 30 (2): 722– 46. https://doi.org/10.1080/09692290.2022.2069145.

Gu, H. 2023. Data, Big Tech, and the New Concept of Sovereignty. *Journal of Chinese Political Science*. https://doi.org/10.1007/s11366-023-09855-1.

He, L. 2024. Nvidia names Huawei a top competitor in major areas including AI chips. https://www.cnn.com/2024/02/23/business/china-nvidia-huawei-competitor-ai-chips-intl-hnk/index.html. Accessed: 2024-05-01.

Heim, L., Fist, T., Egan, J., Huang, S., Zekany, S., Trager, R., Osborne, M. A. and Zilberman, N.. 2024. Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation. arXiv.2403.08501.

Heath, A. 2024. Mark Zuckerberg's New Goal is creating Artificial General Intelligence. https://www.theverge.com/2024/1/18/24042354/mark-zuckerberg-meta-agi-reorg-interview

Herr, T. 2020. *Four Myths about the Cloud: The Geopolitics of Cloud Computing*. Washington, DC: Atlantic Council, Scowcroft Center for Strategy and Security.

Komaitis, K. 2017. The "Wicked Problem" of Data Localisation. *Journal of Cyber Policy* 2 (3): 355–65. https://doi.org/10.1080/23738871.2017.1402942.

Lehdonvirta, V. 2023. Behind AI, a Massive Infrastructure Is Changing Geopolitics. *Oxford Internet Institute.* https://www.oii.ox.ac.uk/news-events/news/behind-ai-amassive-infrastructure-is-changing-geopolitics. Accessed: 2024-05-01.

Lehdonvirta, V.; Wu, B.; and Hawkins, Z. 2023. Cloud Empires' Physical Footprint: How Trade and Security Politics Shape the Global Expansion of U.S. and Chinese Data Centre Infrastructures. *SSRN*. https://doi.org/10.2139/ssrn.4670764.

Lewis, J. A. 2023. An Overview of Global Cloud Competition. *Center for Strategic and International Studies.* https://www.csis.org/analysis/overview-global-cloud-competition. Accessed: 2024-05-01.

Miller, C. 2022. *Chip War: The Fight for the World's Most Critical Technology*. London: Simon & Schuster UK.

Mikler, J. 2018. *The Political Power of Global Corporations.* John Wiley & Sons.

Nanni, R. 2022. Digital Sovereignty and Internet Standards: Normative Implications of Public-Private Relations among Chinese Stakeholders in the Internet Engineering Task Force. *Information, Communication & Society* 25 (16): 2342–62. https://doi.org/10.1080/1369118X.2022.2129270.

Farrell, H. and Newman, A. 2023. *Underground Empire.* Allen Lane.

National AI Research Resource Task Force. 2023. Strengthening the US Artificial Intelligence Innovation Ecosystem. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf. Accessed: 2024-05-01.

Negroponte, N. 1995. *Being digital*. Knopf.

OECD. 2023. OECD Digital Economy Papers: A Blueprint for Building National Compute Capacity for Artificial Intelligence. https://doi.org/10.1787/876367e3-en. Accessed: 2024-05-01.

Pilz, K.; and Heim, L. 2023. Compute at Scale: A Broad Investigation into the Data Center Industry. arXiv.2311.02651.

Png, M. 2022. At the Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. USA: ACM. https://doi.org/10.1145/3531146.3533200.

Pohle, J.; and Thiel, T. 2020. Digital Sovereignty. *Internet Policy Review* 9 (4). https://doi.org/10.14763/2020.4.1532.

Radu, R. 2021. Steering the Governance of Artificial Intelligence: National Strategies in Perspective. *Policy and Society* 40(2): 17893. https://doi.org/10.1080/14494035.2021.1929728.

Radu, R.; and Amon, C. 2021. The Governance of 5G Infrastructure: Between Path Dependency and Risk-Based Approaches. *Journal of Cybersecurity* 7 (1): tyab017. https://doi.org/10.1093/cybsec/tyab017.

Reinsch, W. A.; Schleich, M.; and Denamiel, T. 2023. Insight into the U.S. Semiconductor Export Controls Update. *Center for Strategic and International Studies.* https://www.csis.org/analysis/insight-us-semiconductor-export-controls-update. Accessed: 2024-05-01.

Roberts, R.; Hine, E.; and Floridi, L. 2023. Digital Sovereignty, Digital Expansionism, and the Prospects for Global AI Governance. *SSRN. https://ssrn.com/abstract=4483271.*

Rogers, E. M. 2010. *Diffusion of Innovations*. 4th Edition. London: Simon and Schuster UK.

Rone, J. 2023. The Shape of the Cloud: Contesting Date Centre Construction in North Holland. *New Media & Society*. https://doi.org/10.1177/14614448221145928.

Roth, E. 2023. Microsoft Spent Hundreds of Millions of Dollars on a ChatGPT Supercomputer. https://www.theverge.com/2023/3/13/23637675/microsoft-chatgpt-bing-millions-dollars-supercomputer-openai. Accessed: 2024-05-01.

Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; and O'Keefe, C. 2024. Computing Power and the Governance of Artificial Intelligence'. arXiv.2402.08787.

Sevilla, J.; Heim, L.; Ho, L.; Besiroglu, T.; Hobbhahn, M.; and Villalobos, P. 2022. Compute Trends Across Three Eras of Machine Learning. In *Proceedings 2022 International Joint Conference on Neural Networks*.. https://doi.org/10.1109/IJCNN55064.2022.9891914.

Tang, M. 2022. The Challenge of the Cloud: Between Transnational Capitalism and Data Sovereignty. *Information, Communication & Society* 25(16): 2397-2411. https://doi.org/10.1080/1369118X.2022.2128598

Tardieu, H. 2022. Role of Gaia-X in the European Data Space Ecosystem. In *Designing Data Spaces: The Ecosystem Approach to Competitive Advantage*, edited by B. Otto,; M. ten Hompel,; and S. Wrobel. Springer.

The Economist Intelligence Unit. 2024. China boosts state-led chip investment. https://www.eiu.com/n/china-boosts-state-led-chip-investment. Accessed 2024-05-13.

Trager, R.; Harack, B.; Reuel, A.; Carnegie, A.; Heim, L.; Ho, L.; and Kreps, S. 2023. International Governance of Civilian AI: A Jurisdictional Certification Approach. arXiv:2308.15514.

Vipra, J.; and West, S.M. 2023. Computational Power and AI. *AI Now Institute.* https://ainowinstitute.org/wp-content/uploads/2023/09/AI-Now_Computational-Power-an-AI.pdf. Accessed 2024-05-01.

Wang, Y.; and Chen, D. 2018. Rising Sino-U.S. Competition in Artificial Intelligence. *China Quarterly of International Strategic Studies* 4 (02): 241–58. https://doi.org/10.1142/S2377740018500148.

World Cities Database. 2024. World Cities Database. *Simplemaps: Interactive Maps & Data.* https://simplemaps.com/data/world-cities. Accessed: 2024-05-01.

Yeung, H. W. 2022. Explaining Geographic Shifts of Chip Making toward East Asia and Market Dynamics in Semiconductor Global Production Networks. *Economic Geography* 98 (3): 272–98. https://doi.org/10.1080/00130095.2021.2019010.

Yu, D.; Rosenfeld, H.; and Gupta, A. 2023. The 'AI divide' between the Global North and the Global South. https://www.weforum.org/agenda/2023/01/davos23-ai-divide-global-north-globalsouth/. Accessed: 2024-05-01.